UNIVERSITÉ PARIS-SUD XI – FACULTÉ DE MÉDECINE
École Doctorale ED420 – Santé Publique
Spécialité : Génétique statistique

THÈSE DE DOCTORAT

Discipline : génétique des populations et pharmacogénétique

Différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments

Présentée par

Blandine PATILLON

Encadrement

Audrey SABBAGH et **Emmanuelle GÉNIN**

Jury

Pr Laurent BECQUEMONT	Président du jury
Pr Jean-Louis SERRE	Rapporteur
Dr Lluis QUINTANA-MURCI	Rapporteur
Dr Jérôme CLAIN	Examinateur
Dr Audrey SABBAGH	Directrice de thèse
Dr Emmanuelle GÉNIN	Directrice de thèse

Paris, le 16 juillet 2014

Aux longues ascensions que l'on fait la nuit, en solitaire, par la face nord, en hiver... et sans oxygène.

En démarrant ma thèse de science, il y a quatre ans, j'étais loin de m'attendre à vivre une telle aventure. Dense, tumultueuse, cette expérience hors du commun a littéralement sculpté mon monde durant plusieurs années. Mener à bien cette thèse m'a parfois semblé un but lointain et inaccessible, et à présent qu'il se concrétise, je le regarde profondément heureuse, changée, grandie. Plus patiente aussi.

En dépit de la solitude qui, par définition, caractérise le travail de thèse, jamais le mien n'aurait été possible sans l'intervention de nombreuses personnes que je tiens à remercier aujourd'hui.

Mes premiers remerciements vont à mes deux directrices de thèse, Audrey et Emmanuelle, qui se sont engagées avec moi dans cette longue course de fond, exigeante, difficile, toute en rebondissements. Un grand merci pour votre attention tout au long de ce parcours, la qualité de vos conseils, votre passion pour la recherche, réelle et stimulante et votre investissement lors du sprint final. Vous m'avez appris à réfléchir, à me poser des questions, à analyser, avec l'exigence et la rigueur qu'exige le travail du scientifique. Vous m'avez permis de mener à bien une thèse dont je suis, au final, fière, et qui n'aurait jamais eu cette qualité sans votre regard.
Un remerciement tout spécial pour toi Audrey. Avec ton enthousiasme et ta motivation pour ce projet, tu m'as beaucoup aidée à me dépasser pour poursuivre cette thèse jusqu'au bout. Ta grande disponibilité et ta belle énergie ont été des éléments déterminants pour moi, et sans ton exemple je ne serai pas là aujourd'hui. La grande question désormais va être d'apprendre à travailler sans toi !

Je remercie chaleureusement les membres du jury, Jean-Louis Serre et Lluis Quintana-Murci pour leur lecture attentive de ce manuscrit sur lequel ils ont rédigé un rapport, ainsi que Laurent Becquemont et Jérôme Clain pour avoir accepté de faire partie du jury de soutenance.

De tout cœur, je remercie Jean Bouyer, le directeur de l'École Doctorale 420 et Audrey Bourgeois son assistante, qui avec beaucoup de disponibilité, de compréhension et de gentillesse ont adouci toutes sortes de difficultés et m'ont encouragée à poursuivre jusqu'au bout ce travail.
Je remercie sincèrement Nadia Iafrate, du service du personnel, pour son aide efficace qui m'a bien soulagée au moment délicat de mon opération.
Je remercie également l'Université Paris-Sud XI qui m'a soutenue financièrement le plus longtemps possible.

Je tiens à remercier Florence Demenais et Philippe Deloron pour m'avoir accueillie chacun dans leur laboratoire et fourni un cadre propice pour mener à bien ma thèse.
Je salue plus généralement tous les membres de l'unité INSERM UMR 946 et du CEPH qui m'ont entourée les premières années : Li, Mourad, Anne-Louise, Martine, Patricia Emmanuelle, Marie-Hélène. Une mention spéciale à Hamida, pour son aide précieuse dans la résolution des situations informatiques les plus tragiques.
Je salue chaleureusement toute l'équipe IRD UMR 216 dont la remarquable joie de vivre et le dynamisme m'ont permis de travailler dans une ambiance vivante et motivante. En particulier, les membres du pigeonnier : Jean Gérard, Brigitte, André, Jean-Yves, Jean-Christophe, Alexandre, Gilles. Et aussi Pascale !

Surtout, surtout, je voudrais remercier tous mes camarades doctorants, les seuls êtres au monde capables de comprendre véritablement ce qu'a été mon quotidien pendant quatre ans (à défaut de comprendre mon sujet de recherche). Un immense merci donc

au grand, fort et intelligent, le plus impensable de mes amis, Steven. Tu vois, malgré l'absence notable il est vrai, de cerveau d'éléphant formolé sur notre bureau, d'un taux plasmatique minimal de vitamine D, mais aussi de résultat de recherche époustouflant permettant en potentiel de guérir l'intégralité de l'humanité, on s'en sort bien finalement, d'une thèse en génétique statistique !

à Violeta, ma géniale, curieuse, intense, vibrante amie ! Quelle passion, quelle énergie, quel enthousiasme, qui vont t'emmener loin !

à Pierre, mon compagnon de travail, partenaire de tous mes articles, génie de la génétique des populations, inventeur des meilleurs paramètres des tests de sélection sur les données de séquence,

à Myriam, ma courageuse, volontaire, battante, admirable Myriam, qui a partagé avec une belle amitié toutes les péripéties de ma thèse,

à Manfred, dont la présence calme, réconfortante dans mon bureau, la grande bienveillance m'ont tellement aidée à regagner la confiance qui m'avait quittée,

A Sébastien, le spécialiste du génome humain et des mystères de son annotation, mais aussi du *Perl* et des réponses à mes questions insolubles,

à Tania, le seul spécimen de doctorant qui soit, à ma connaissance, capable de fonctionner en quasi-exclusivité à la photosynthèse. Si c'est cela qui te rend aussi adorable et ravie en permanence, s'il te plaît, apprends-moi !

à Rémi qui n'est plus en PhD aujourd'hui, mais qui m'a prodigué soutien, encouragements et conseils utiles tout au long de ma thèse,

A Géraud, toujours calme, compréhensif, perspicace, qui va rentrer avec bonheur dans l'après-thèse presque en même temps que moi,

à Gaëlle, pour le charme de nos pauses thé, l'intensité de nos soirées codage en R et tes sages conseils sur la manière d'être un bon doctorant,

Merci aussi à Julie, Chloé, Laure, Amaury, Nolwenn, pour ce quotidien partagé, à travers des conseils techniques, des repas, des séminaires…

Je pense aussi à Hervé, loin, loin là-bas ! Merci pour ta présence, tes conseils en statistiques, en choix de bière, en musique, en tout d'ailleurs, pour ton amitié.

Même s'il n'est plus avec nous, ou peut-être d'une autre manière, je pense très fort à Howard, mon grand-père académique qui m'a soutenue, encouragée, valorisée avec tant de gentillesse et d'humour. Howard m'a appris à rédiger une phrase d'article efficace et simple, et a contribué avec énergie à la passionnante élaboration de la meilleure stratégie d'analyse des signaux d'XP-CLR permettant de délimiter au plus précisément la région génomique exacte soumise à sélection en Asie de l'Est autour du gène *VKORC1*.

Je suis heureuse aussi de saluer l'équipe du Département Santé Publique et Biostatistique de la Faculté de Pharmacie : Emmanuel, Simone, Chantal, Virginie, Ioannis, Patrick, Mohammed, avec qui j'ai appris à préparer des TDs en statistiques et informatique et à encadrer des étudiants ; expérience très motivante dont je garde un excellent souvenir.

Je t'embrasse toi aussi ma petite Constance, ma brillante élève qui va devenir une future grande statisticienne, qui se bat comme un lion avec un courage admirable.

Je remercie vivement mes parents, mes sœurs Marie, Hélène et Sophie, et mes chers grands-parents qui m'ont tous, chacun à leur manière accompagnée avec attention et amour. Si manifestement aucun d'eux, aujourd'hui encore, ne comprend – même approximativement – le titre de ma thèse, ils savent et comprennent mieux que quiconque ce que ce travail a représenté pour moi. Je sais la chance que j'ai de les avoir autour de moi et les embrasse de tout mon cœur. Je remercie tout particulièrement mes parents, pour m'avoir transmis le goût de l'effort et du plein engagement dans mes projets. Forte de cela, je me sens aujourd'hui capable d'aller audacieusement m'épanouir dans une vie où l'on ne nous distribue pas de carte d'étudiant.

À mes amis ! Que ferais-je, que serais-je sans vous ! Je remercie surtout, du fond du cœur Elisabeth (R !), Pauline et Elisabeth (L !) pour leur immense et infaillible soutien, qui m'a énormément aidée à ouvrir l'espace et prendre du recul quand mon monde se resserrait trop autour de ma thèse.
J'embrasse très fort Claire et Camille, qui m'ont fourni avantageusement gîte et couvert à certains instants cruciaux de la rédaction, et qui me font tellement, tellement rire.
Une pensée affectueuse à mon cher cousin Julien, à Marie-Camille, Cécile, Clémentine, Thibaut, Mijo, Mouch, Anne-Sophie, Aline, Sophie et ma marraine Blandine qui apportent tant de couleurs et de gaieté dans ma vie. Et puis à tous les amis qui me suivent de plus loin, de façon plus épisodique, qui comptent pourtant tout autant.

Enfin, un immense merci à toi Florian, mon plus bel allié, qui accompagne avec douceur et patience tous les mouvements de ma vie. Tu sais Paul McCartney c'est trompé, c'est toi la plus belle chanson du monde. Quel bonheur c'est pour moi de l'écouter chaque jour.

UMR 216 IRD « Mère et enfant face aux infections tropicales »

Université Paris Descartes, Faculté des Sciences Pharmaceutiques et Biologiques, Laboratoire de parasitologie

4, Avenue de l'Observatoire

75270 Paris Cedex 6

01 70 64 94 29

audrey.sabbagh@ird.fr

blandine.patillon@gmail.com

Inserm U1078 « Génétique, Génomique fonctionnelle et Biotechnologies »

46, rue Félix Le Dantec

CS 51819

29218 Brest Cedex 2

02 98 22 34 08

emmanuellle.genin@inserm.fr

Titre

Différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments.

Résumé

Tous les individus ne répondent pas de la même façon à un même traitement médicamenteux, tant sur le plan pharmacologique (efficacité) que sur le plan toxicologique (effets indésirables). Des facteurs génétiques affectant la pharmacocinétique et la pharmacodynamie des médicaments jouent un rôle déterminant dans cette variabilité interindividuelle de réponse. Certains de ces facteurs sont distribués de manière hétérogène entre les populations humaines. Ces différences s'expliquent en partie par des phénomènes d'adaptation locale des populations à leur environnement. Au cours de son histoire, l'homme a dû en effet faire face à des changements de son environnement chimique, qui ont entraîné des pressions de sélection naturelle sur les gènes intervenant dans la réponse de l'organisme aux xénobiotiques. Ce sont ces mêmes gènes qui, aujourd'hui, influencent la réponse aux médicaments.

La formidable accélération des progrès de la génétique donne accès aujourd'hui à la variabilité génétique des populations humaines sur l'ensemble du génome, facilitant la découverte et la compréhension des mécanismes génétiques à l'origine des traits complexes comme la réponse aux médicaments. Les outils de la génétique des populations permettent notamment d'identifier des variants affichant un niveau de différenciation génétique inhabituel entre les populations humaines et de déterminer dans quelle mesure la sélection naturelle a joué un rôle dans les profils atypiques observés.

Dans cette thèse, nous avons appliqué ces outils à des données de génotypage et de séquençage pour analyser les profils de différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments. Nous avons ainsi démontré qu'une sélection positive récente en Asie de l'Est dans la région génomique du gène *VKORC1* était responsable d'une hétérogénéité de distribution du variant fonctionnel de *VKORC1*, à l'origine des différences de sensibilité génétique aux

anticoagulant oraux de type antivitamine K entre les populations humaines. Puis, en étendant notre analyse à l'ensemble des pharmacogènes majeurs, nous avons identifié de nouveaux variants potentiellement intéressants en pharmacogénétique pour expliquer les différences de réponse aux médicaments entre les populations humaines et les individus. Enfin, l'étude approfondie du gène *NAT2* nous a permis de révéler un processus de sélection homogénéisante ciblant un variant fonctionnel associé à un phénotype d'acétylation très lent. Ces résultats soulignent l'influence déterminante de la sélection naturelle dans la variabilité de réponse aux médicaments entre les populations et les individus. Ils montrent l'apport de la génétique des populations pour une meilleure compréhension de la composante génétique de la réponse aux médicaments et des traits complexes.

Mots clés

Génétique des populations – sélection naturelle – balayage sélectif – différenciation génétique – réponse aux médicaments – pharmacogénétique – Projet 1000 Génomes – Panel HGDP-CEPH – antivitamine K – *VKORC1* – *NAT2* – pharmacogènes.

Title

Genetic differentiation of human populations for genes involved in drug response.

Abstract

Response to drug treatment can be highly variable between individuals, both in terms of therapeutic effect (efficacy) and of adverse reactions (toxicity). Genetic factors affecting drug pharmacodynamics and pharmacokinetics play a major role in this inter-individual variability. Some of these factors are heterogeneously distributed among human populations. Local adaptation of populations to their environment partly explained those differences. Indeed, during human evolution, populations had to cope with changes in their chemical environment that triggered selective pressures on genes involved in xenobiotic response. Those genes are the same ones that influence drug response today.

The tremendous recent advances in genotyping and sequencing technologies now provide access to the genome-wide patterns of genetic variation in a growing number of human populations, facilitating our understanding of the genetic mechanisms underlying complex traits such as drug response. Population genetic tools allow the identification of variants showing an unusual pattern of genetic differentiation among human populations and the determination of the role played by natural selection in shaping the atypical patterns observed.

In this thesis, we have applied these tools on both SNP-chip genotyping data and Next Generation Sequencing data to analyze the genetic differentiation patterns of human populations for genes involved in drug response. We show that a nearly complete selective sweep in East Asia in the genomic region of the *VKORC1* gene is responsible for an heterogeneous distribution of the *VKORC1* functional variant and can explain the inter-population genetic differences in response to oral anti-vitamin K anticoagulants. Extending the analysis to all major pharmacogenes, we have identified new variants of potential relevance to pharmacogenetics which could explain inter-

population and inter-individual differences in drug response. Finally, by a comprehensive analysis of the *NAT2* gene, we evidence a homogenizing selection process targeting a functional variant associated with a very slow acetylation phenotype. These results emphasize the crucial role of natural selection in the inter-population and inter-individual drug response variability. They also illustrate the relevance of population genetics studies for a better understanding of the genetic component underlying drug response and complex traits.

Key words

Population genetics – natural selection – selective sweep – genetic differentiation – drug response – pharmacogenetics – 1000 Genomes Project – HGDP-CEPH Panel – antivitamin K – VKORC1 – NAT2 – pharmacogene.

Articles publiés

Patillon B, Luisi P, Blanché H, Patin E, Cann HM, Génin E, Sabbagh A. *Positive selection in the chromosome 16 VKORC1 genomic region has contributed to the variability of anticoagulant response in humans*. PLoS One. 2012;7(12):e53049. PMID: 23285254

Patillon B, Génin E, Sabbagh A. *La variabilité de réponse aux anticoagulants oraux, une conséquence de la sélection naturelle*. Médecine Sciences, 2013 ;29:159.

Articles en préparation (2)

Patillon B, Luisi P, Poloni ES, Boukouvala S, Darlu P, Génin E, Sabbagh A. *A homogenizing process of selection has maintained an 'ultra-slow' acetylation NAT2 variant in humans*. Soumis à Human Biology.

Patillon B, Luisi P, Letort S, Laayouni H, Génin E, Sabbagh A. *Global patterns of population genetic differentiation for genes involved in drug response*. En cours de rédaction.

Article résultant de travaux additionnels pendant la thèse (1)

Gineau L, Luisi P, Castelli E, Milet J, Courtin D, **Patillon B**, Laayouni H, Moreau P, Donadi EA, Garcia A, Sabbagh A. *Balancing immunity and tolerance: Genetic footprint of natural selection at the HLA-G promoter region*. Soumis à *Human Genetics.*

Communications orales (2)

Patillon B, Génin E, Sabbagh A. Genetic differentiation of human populations for drug response. 7e Assises de Génétique Humaine et Médicale, Bordeaux, France, 29-31Janvier 2014.

Patillon B, Génin E, Sabbagh A. *Différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments.* 31e Colloque du Groupement des Anthropologistes de Langue Française (GALF), Marseille, France, 16-18 Octobre 2013.

Communications par affiche (3)

Patillon B, Luisi P, Génin E, Sabbagh A. *Worldwide population differentiation of genes involved in drug response*. 4e Journée scientifique de l'Institut Médicament Toxicologie Chimie Environnement (IMTCE), Paris, France, 31 May 2013.

Patillon B, Luisi P, Génin E, Sabbagh A. *Impact of natural selection on genes involved in drug response*. 62e Réunion annuelle de l'European Society of Human Genetics (ESHG), Paris, France, June 8-11, 2013.

Patillon B, Luisi P, Blanché H, Patin E, Cann HM, Génin E, Sabbagh A. *Detecting signatures of recent positive selection at the VKORC1 gene locus*. 20e Réunion annuelle de l'International Genetic Epidemiology Society (IGES), Heidelberg, Germany, 18-20 September 2011.

ADME	Absorption, Distribution, Métabolisme et Élimination
ANSM	Agence nationale de sécurité du médicament et des produits de santé
AVK	Antivitamine K
CEPH	Centre d'Étude du Polymorphisme Humain
CPIC	*Clinical Pharmacogenetics Implementation Consortium*
CYP	Cytochrome P450
DL	Déséquilibre de liaison
EHH	*Extended haplotype homozygosity*
FDA	*Food and Drug Administration*
GWAS	*Genome wide association study*
GWSS	*Genome wide selection scan*
HAS	Haute Autorité de Santé
HSA	*Health Sciences Authority* (Singapour)
HGDP	*Human Genome Diversity Project*
iHS	*Integrated standardized EHH*
INR	*International Normalized Ratio*
IWPC	*International Warfarin Pharmacogenetics Consortium*
LRH	*Long range haplotype*
MAF	*Minor allele frequency*
NAT2	N-acétyltransférase 2
NCBI	*National Center for Biotechnology Information*
NIH	*National Institutes of Health*
OMS	Organisation Mondiale de la Santé
Pb	Paire de base
PGRN	*Pharmacogenomics Research Network*
PharmGKB	*Pharmacogenomics Knowledge Base*
SNP	*Single nucleotide polymorphism*

UTR	*Untranslated Region*
VIP	*Very Important Pharmacogene*
VKORC1	Sous-unité 1 du complexe vitamine K époxyde réductase
XP-CLR	*Cross population composite likelihood ratio method*
XP-EHH	*Cross population extended haplotype homozygosity*

Introduction

La réponse aux médicaments est extrêmement variable entre les individus, que ce soit en terme d'efficacité de traitement (absence de réponse pharmacologique), ou de toxicité (survenue d'effets indésirables qui peuvent être graves voire mortels). Cette variabilité interindividuelle est chaque année responsable d'un grand nombre d'accidents iatrogéniques, qui représentent un problème de santé publique majeur. Depuis la découverte de l'existence de facteurs génétiques intervenant dans la réponse aux médicaments il y a une soixantaine d'années, le champ de la pharmacogénétique s'est développé de façon spectaculaire et a permis de révéler l'importance des déterminants génétiques dans la variabilité de la réponse aux médicaments. A l'heure actuelle, l'une des problématiques majeures de la pharmacogénétique réside dans l'identification précise des variants pharmacogénétiques qui pourraient être utilisés en pratique médicale en tant que biomarqueurs des médicaments, favorisant ainsi l'optimisation des stratégies thérapeutiques au profil génétique du patient.

Les médicaments font partie des xénobiotiques, substances chimiques étrangères à l'organisme, auxquels l'homme est exposé en permanence. Durant leur histoire évolutive, les populations humaines ont dû faire face à des variations de leur environnement chimique, liées par exemple à des modifications d'ordre climatique ou alimentaire. L'adaptation des populations à ces nouveaux environnements a été permise par l'action de la sélection naturelle sur les gènes impliqués dans la réponse aux xénobiotiques. Le plus souvent, ces phénomènes adaptatifs se sont passés à une échelle locale, c'est-à-dire qu'ils ont concerné uniquement les populations soumises au changement environnemental. Il en résulte une différenciation géographique importante à l'échelle du globe pour les gènes gouvernant la réponse aux xénobiotiques et aux médicaments. Aujourd'hui, cette histoire évolutive différentielle des populations se reflète par une variabilité phénotypique de réponse aux médicaments entre populations.

Sur le génome, des empreintes de la sélection naturelle sont visibles au niveau des régions génomiques hébergeant les gènes de la réponse aux médicaments. Les progrès réalisés ces dernières années dans les techniques

de caractérisation du génome humain, en combinaison avec le développement des outils de la génétique des populations, rendent possible leur détection. L'étude de la différenciation génétique des populations et des pressions sélectives sous-jacentes permet d'apporter des clés majeures dans la compréhension de la variabilité interindividuelle et inter-populationnelle dans la susceptibilité génétique aux traits complexes comme la réponse aux médicaments.

Dans cette thèse, nous mettons à profit différents jeux de données génétiques dans différentes populations pour étudier les profils de différenciation génétique des populations humaines pour les gènes majeurs de la réponse aux médicaments. Trois objectifs principaux ont orienté ce travail : (1) évaluer comment ces profils diffèrent en fonction de la catégorie pharmacogénétique à laquelle appartiennent ces gènes ; (2) sélectionner les variants pharmacogénétiques présentant un profil de différenciation génétique extrême et évaluer dans quelle mesure ces profils sont le résultat de l'action de la sélection naturelle et (3) identifier des nouveaux variants potentiellement intéressants en pharmacogénétique, qui seraient susceptibles d'expliquer une partie de la variabilité observée dans la réponse clinique aux médicaments entre les populations et les individus.

Cette thèse est organisée en quatre parties. La première introduit les grands principes de la pharmacogénétique et décrit les méthodes de génétique des populations permettant de quantifier la différenciation génétique inter-populationnelle et de détecter les signatures de sélection naturelle sur le génome. La deuxième partie présente l'application de ces méthodes au gène *VKORC1*, codant pour la cible pharmacologique directe des anticoagulants de type antivitamine K (AVK), dans les données de génotypage du Panel HGDP-CEPH (*Human Genome Diversity Project*). Nous y verrons qu'un phénomène de sélection positive récent est sans doute responsable du profil de différenciation atypique du variant fonctionnel rs9923231 de *VKORC1*, qui confère un phénotype de sensibilité augmentée aux AVK. Cet événement sélectif se traduit aujourd'hui par des différences significatives de sensibilité génétique aux AVK entre les populations humaines.

Dans la troisième partie, nous étendons cette approche à l'ensemble des pharmacogènes les plus importants en étudiant leur profil de différenciation génétique dans les données de séquençage du Projet 1000 Génomes. Nous identifions des nouveaux variants d'intérêt en pharmacogénétique présentant à la fois un profil de différenciation extrême ainsi qu'une forte probabilité d'être délétères, telle qu'estimée par des outils bioinformatiques de prédiction de l'effet fonctionnel. Nous évaluons également le rôle de la sélection positive dans la détermination de ces profils de différenciation atypiques. La quatrième partie se concentre sur l'étude du gène *NAT2* impliqué dans les mécanismes de détoxication et d'élimination de nombreux médicaments importants en clinique. Sur la base des données de séquence du projet 1000 Génomes, nous mettons en évidence un niveau de différenciation génétique inhabituellement faible pour le variant fonctionnel rs1799930, associé à un phénotype d'acétylation très lent, suggérant qu'un processus de sélection homogénéisante a été à l'œuvre dans cette région du génome. Nous terminons par une discussion exposant les difficultés rencontrées par la pharmacogénétique pour son intégration dans la pratique clinique et discutons de l'apport de la génétique des populations à l'identification des facteurs génétiques impliqués dans la réponse aux médicaments. Les limites de nos études, notamment notre capacité à détecter les signatures génomiques de la sélection, sont également discutées.

Partie 1

Introduction à la pharmacogénétique et à la génétique des populations

Chapitre 1

Variabilité de réponse aux médicaments et pharmacogénétique

Les individus ne sont pas égaux face aux médicaments, et peuvent présenter des réactions extrêmement variables en réponse à un même traitement médicamenteux, en dépit de conditions d'administration identiques. Alors que certains vont s'éloigner de la réponse attendue en présentant une diminution ou une absence d'efficacité ; d'autres vont développer des réactions indésirables voire des toxicités importantes qui peuvent être très dangereuses pour la santé de l'individu. Si cette variabilité interindividuelle de réponse peut être liée à des facteurs environnementaux ou à des états physiopathologiques particuliers, les facteurs génétiques semblent jouer un rôle important. L'objectif de la pharmacogénétique est d'identifier les individus présentant un risque particulier d'inefficacité ou de toxicité vis-à-vis de certains médicaments sur la base de leurs profils génétiques afin d'optimiser les traitements médicamenteux, tant en termes d'efficacité que de sécurité d'emploi.

1. La variabilité de réponse aux médicaments : conséquences sanitaires et socio-économiques

La variabilité de réponse aux traitements médicamenteux, souvent difficile à prévoir, constitue une limitation importante à l'emploi des médicaments. Elle pose un réel problème de santé publique et a un impact considérable sur le plan économique et social.

1.1 Conséquences sanitaires

Les effets indésirables des médicaments, définis par des réactions non souhaitées et nocives pour la santé survenant suite à l'administration d'un médicament à des doses standards (World Health Organization, 1969), représentent une part importante de la iatrogénie médicamenteuse, et sont à l'origine d'un véritable problème de santé publique au niveau mondial.

En terme de toxicité

Dans les années 1990, différentes études menées aux États-Unis ont estimé que les effets indésirables des médicaments, étaient à l'origine de 2,4 à 6,5 % des hospitalisations (Bates et al., 1995; Classen et al., 1997; Leape et al., 1991). Un travail réalisé en 1998, a montré qu'à eux seuls, les effets indésirables graves étaient responsables de plus de deux millions d'hospitalisations et de 100 000 décès par an aux États-Unis, les classant entre la quatrième et la sixième cause de mortalité (Lazarou et al., 1998).

En France, les chiffres d'une enquête menée en 1998 par le réseau des Centres Régionaux de Pharmacovigilance (CRPV) rapportent un taux d'incidence de 3,2 % d'évènements indésirables dus à des médicaments nécessitant une hospitalisation (Pouyanne et al., 2000).

Des études plus récentes, réalisées un peu partout dans le monde, font état de chiffres largement en hausse (Phillips et al., 2014; Pirmohamed et al., 2004). Au Royaume-Uni, une analyse prospective menée sur une durée de six mois en 2004 dans deux grands hôpitaux a montré que les effets indésirables étaient responsables de 6,5 % des hospitalisations (Pirmohamed et al., 2004). Une revue de la littérature basée sur les résultats de 46 études réalisées dans différents pays rapporte une incidence moyenne de 6,1 % d'effets indésirables des médicaments chez les patients hospitalisés (Krähenbühl-Melcher et al., 2007). En France, les chiffres les plus couramment avancés font état de 140 000 hospitalisations provoquées par les accidents médicamenteux et de 18 000 décès avérés annuels, sans compter les accidents bénins qui ne font pas l'objet d'une déclaration systématique[1]. On estime en effet que durant cette dernière décennie, le nombre

[1] Cependant il faut souligner le manque aujourd'hui en France d'étude récente évaluant

d'hospitalisations dues aux effets indésirables des médicaments a augmenté en Angleterre de 76,8 % (Wu et al., 2010). Aux États-Unis entre 1998 et 2005, le nombre d'effets indésirables graves a augmenté d'un facteur 2,6, soit quatre fois plus rapidement que les prescriptions aux USA durant la même période (Moore et al., 2007). Si ces chiffres sont probablement reliés en partie à une amélioration de la déclaration des effets indésirables dans la communauté médicale, ils ont également pour cause le vieillissement de la population (Majeed and Aylin, 2005; McLean and Le Couteur, 2004). L'incidence et la gravité des accidents iatrogéniques augmentent avec l'âge : on estime que ces accidents représentent environ 20 % des hospitalisation des sujets de plus de 80 ans (Roughead et al., 1998) et que les effets indésirables létaux sont plus fréquents chez les patients âgés (Ebbesen et al., 2001). En effet, en plus d'une plus grande fragilité physiologique, cette population est plus susceptible de rencontrer des effets indésirables dus aux médicaments, en raison d'une polymédication quasi-systématique (Mannesse et al., 2000; McLean and Le Couteur, 2004; Petrovic et al., 2012; Wu et al., 2010).

Aujourd'hui, la mortalité et la morbidité liées aux effets indésirables des médicaments sont presque équivalentes à celles du cancer ou des maladies cardiovasculaires dans les pays occidentaux (Duran-Frigola and Aloy, 2013). En France, c'est un fléau qui tue plus que les suicides et les accidents de la route réunis.

En terme d'efficacité thérapeutique

La variabilité de réponse aux médicaments entraîne également des situations d'inefficacité thérapeutique, plus difficiles à appréhender et encore moins évaluées à l'heure actuelle. On estime globalement que pour les médicaments majoritairement utilisés dans le traitement de la plupart des maladies communes, le taux de réponse varie de 25 à 80 % (Spear et al., 2001). Ainsi, on observe par exemple que des traitements très répandus tels que les analgésiques ne permettent pas de traiter toute la population uniformément ; que 35 % des patients ne répondent pas aux béta-bloquants ; et que, dans le cas des maladies mentales, un malade sur deux ne tire aucun

profit de son traitement antidépresseur, et moins d'un patient schizophrène sur trois bénéficie de son traitement antipsychotique.

1.2 Conséquences socio-économiques

Les conséquences économiques liées de façon directe ou indirecte à la toxicité médicamenteuse sont absolument pharamineuses.

Les seuls coûts annuels engendrés par les hospitalisations dues aux effets indésirables des médicaments ont été évalués en 1997 aux États-Unis à près de 80 milliards de dollars (Bates et al., 1995). Une grande enquête prospective réalisée dans deux grands hôpitaux au Royaume-Uni estime que le coût annuel des effets indésirables des médicaments supporté par le système de santé anglais *National Health Service* NHS s'élève à 706 millions d'euros (Pirmohamed et al., 2004). La Commission européenne chiffre ce coût en 2008 à 79 milliards d'euros en Europe. En France, il a été évalué à 3,4 millions d'euros pour un seul hôpital de 339 lits (Lagnaoui et al., 2000). Une étude suédoise très récente reporte un montant annuel de 21 millions de dollars pour 100 000 habitants, ce qui représente approximativement 10 % des dépenses de santé totales (Gyllensten et al., 2014).

Ce coût varie considérablement selon le type d'événement indésirable : il est 600 fois plus élevé pour un effet indésirable qui entraîne des séquelles que pour un effet indésirable non grave.

Globalement, on estime que les coûts liés aux accidents iatrogéniques sont supérieurs à ceux des traitements médicamenteux en eux-mêmes (Sim and Ingelman-Sundberg, 2011). Cette situation est d'autant plus préoccupante qu'elle est amplifiée du fait de l'augmentation drastique de la consommation de médicaments actuellement. En Australie, elle a augmenté cette dernière décennie de 40 % (Phillips et al., 2014). En France, le total des ventes de médicaments est passé de 16,5 à 27,5 milliards d'euros entre 2000 et 2010.

Par ailleurs, les effets indésirables des médicaments posent également un réel problème à l'industrie pharmaceutique, d'une part parce qu'ils sont

responsables de 17 % des arrêt prématurés du développement de nouvelles molécules candidates (Dimasi, 2001) ; d'autre part, parce que leur faible incidence les rend difficilement détectables lors des essais cliniques, ce qui peut conduire au retrait du marché de médicaments en post AMM (Autorisation de Mise sur le Marché) (Uetrecht and Naisbitt, 2013).

La iatrogénie médicamenteuse a également des conséquences sociales importantes pour le malade comme pour la collectivité. Cela se traduit notamment par un grand nombre d'arrêts de travail et d'interruption d'activités (43 % d'arrêts chez les actifs) (Apretna et al., 2005). Le retentissement de la iatrogénie sur les activités quotidiennes et la qualité de vie du patient est parfois tel qu'il peut induire une méfiance, voire un rejet, vis à vis des traitements médicamenteux (Wu et al., 2010). Dans le monde, et particulièrement en France, les crises sanitaires récentes liées aux effets indésirables des médicaments (Médiator®, contraceptifs oraux, vaccins contre l'hépatite B, psychotropes, etc.) ont terni dans son ensemble l'image du médicament : il est désormais vécu par beaucoup, plus comme un risque à éviter que comme un outil majeur de santé publique.

L'impact humain, psychologique, sanitaire, financier et social de la iatrogénie médicamenteuse est, de toute évidence, considérable ; tant par le nombre de maladies et de complications qui pourraient être évitées par des traitements médicamenteux davantage optimisés à l'échelle individuelle que par les coûts supportés par la collectivité.

1.3 Concept de la médecine personnalisée

Considérant l'ampleur des conséquences liées à la variabilité de réponse aux médicaments, il apparaît nécessaire d'augmenter la connaissance des facteurs prédictifs de la réponse aux médicaments, aussi bien pour le patient, afin d'améliorer sa prise en charge thérapeutique, que pour la collectivité, afin de réduire les coûts d'hospitalisation liés aux effets indésirables et aux inefficacités thérapeutiques.

En tenant compte de l'ensemble de ces facteurs, il pourrait alors être possible d'adapter les traitements médicamenteux à chaque patient. Cette stratégie de choix a donné naissance, ces dernières années, au concept de la médecine personnalisée (Encadré 1.1).

Encadré 1.1 I La médecine personnalisée

Bien que la médecine ait toujours été, de façon évidente, pratiquée à l'échelle individuelle en tenant compte pour chaque patient de ses facteurs physiopathologiques, son histoire familiale, son mode de vie, etc., le concept de médecine personnalisée a pris ces dernières années une connotation nouvelle, fondée sur la connaissance des bases moléculaires des maladies.

La médecine personnalisée consiste à choisir le traitement le plus adapté en fonction du profil biologique du patient et des caractéristiques moléculaires de sa maladie. Son objectif est résumé par la phrase désormais fameuse : « le bon médicament, à la bon dose, au bon patient, au bon moment », énoncée par Margaret Hamburg, commissaire à la FDA (*Food and Drug Administration*) en 2010.

L'amélioration de la mise en place d'une médecine personnalisée dans la pratique clinique représente, progressivement, l'un des plus grands enjeux de la médecine de demain.
Les principaux axes de recherche dans ce domaine sont (Scott, 2011) :

- l'identification précise des facteurs génétiques impliqués dans les maladies complexes,
- l'étude des mécanismes d'interactions gène-environnement qui sont à l'origine de l'apparition des maladies,
- l'utilisation de biomarqueurs pharmacogénétiques permettant d'augmenter l'efficacité et de diminuer la toxicité des thérapies médicamenteuses.

Le concept de la médecine personnalisée constituait l'une des aspirations déclarées du Projet génome humain (Meyer et al., 2012). En 2010, il est à l'origine de la mise en place d'une collaboration entre la FDA et le NIH (*National Institutes of Health*) dont le but est d'accélérer et de favoriser son implémentation en clinique, *via* l'optimisation des prescriptions médicamenteuses et le développement de nouveaux médicaments (Hamburg and Collins, 2010).

2. Sources de variabilité de la réponse aux médicaments

La variabilité interindividuelle des effets pharmacologiques des médicaments, en termes d'efficacité et de tolérance, est un phénomène complexe, se rapportant à de multiples causes.

2.1 Rappel sur les composantes de la réponse aux médicaments

L'effet thérapeutique du médicament est déterminé par deux grandes catégories de réactions : celles de la pharmacodynamie (PD), liées au mode d'action du médicament sur l'organisme ; et celles de la pharmacocinétique (PK), liées à l'inverse au devenir du médicament dans l'organisme.

La pharmacodynamie

Elle correspond à l'ensemble des effets du médicament sur l'organisme. Elle est liée aux interactions entre la substance active et ses cibles d'action (récepteur, enzyme), qui peuvent être plus ou moins bien identifiées sur le plan moléculaire. La pharmacodynamie permet de décrire la relation entre la concentration au site d'action du médicament et la réponse thérapeutique observée.

La pharmacocinétique

A l'inverse, les réactions de la pharmacocinétique concernent l'ensemble des mécanismes de prise en charge du médicament par l'organisme. Elle décrit la relation entre la dose administrée et les concentrations en médicament mesurées. Elles sont composées de quatre grandes étapes cinétiques résumées par l'acronyme international « ADME » :

- L'*absorption* (A), qui aboutit à la pénétration de la substance active dans la circulation systémique, après les étapes de résorption (passage membranaire) et de premiers passages, notamment hépatique, moments de biotransformations métaboliques.
- La *distribution* (D), qui correspond à la répartition de la substance active dans l'organisme à partir de la circulation générale. Elle

dépend fortement du taux de liaison de cette dernière aux protéines plasmatiques.

- Le *métabolisme* (M), qui consiste en la transformation du médicament par l'équipement enzymatique de l'organisme, en particulier au niveau du foie, générant des métabolites plus hydrophiles et plus facilement éliminables, qui peuvent avoir une activité pharmacologique plus ou moins importante (Encadré 1.2).
- L'*élimination* (E) du principe actif et de ses métabolites de l'organisme, qui passe par l'excrétion directe, principalement par le rein, ou après biotransformation (métabolisation).

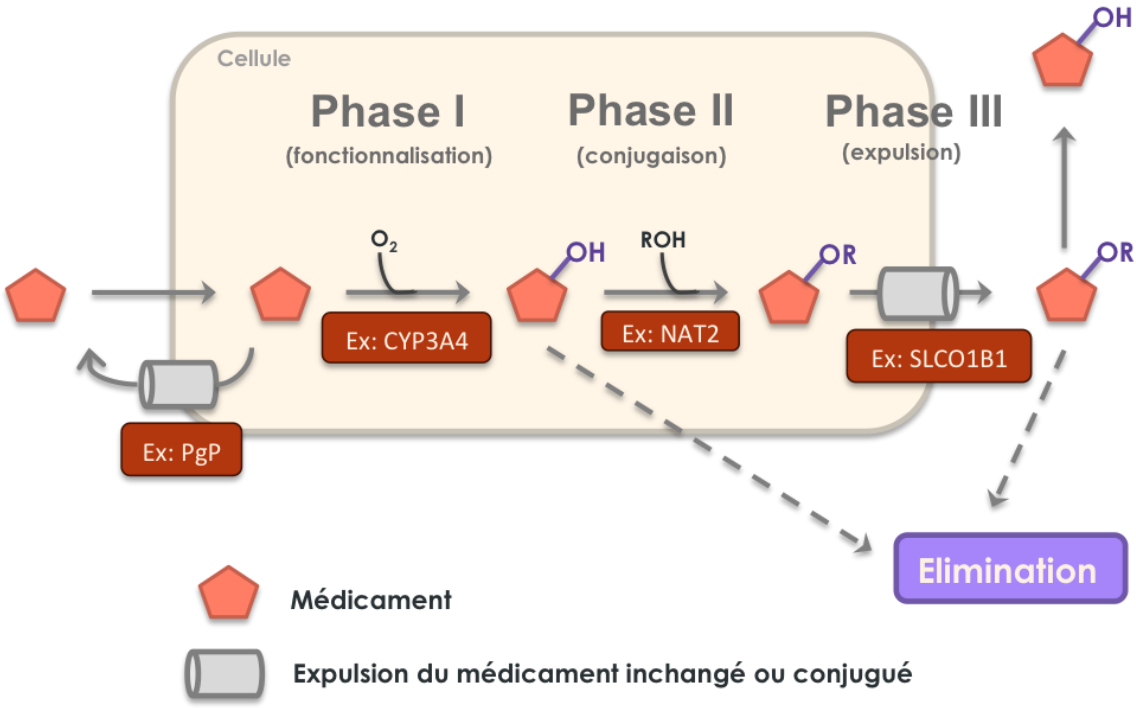


Figure 1.1| Représentation schématique du métabolisme des médicaments (Encadré 1.2).

Encadré 1.2 I Le métabolisme des médicaments

Pour se défendre contre l'agression de substances chimiques potentiellement toxiques, auxquelles il est exposé en permanence, l'organisme humain a mis en place une réponse cellulaire complexe qui permet l'élimination des composés étrangers et qui participe aux processus de détoxification cellulaire. Les médicaments, qui représentent une part non négligeable des xénobiotiques, sont en général de nature hydrophobe, et de ce fait, passent facilement la membrane plasmique. L'entrée d'un médicament dans la cellule va être détectée par les récepteurs des xénobiotiques, qui vont à leur tour induire un ensemble d'activités enzymatiques et de transporteurs permettant leur détoxication et leur élimination.

Le métabolisme d'un médicament correspond à sa transformation enzymatique en un ou plusieurs composés hydrosolubles plus facilement éliminables : les métabolites, qui peuvent être inactifs (directement éliminés), actifs, ou bien parfois toxiques. La biotransformation des médicaments a essentiellement lieu dans le foie, mais également au niveau des reins, poumons, intestins, peau, …

Trois grands systèmes enzymatiques spécialisés et complémentaires interviennent dans cette biotransformation, à travers :

- Les **réactions de phase I** qui conduisent à des dérivés possédant des groupements fonctionnels les rendant plus polaires. Elles comprennent les réactions :
    - d'oxydation qui ont principalement lieu dans le foie, et nécessitent la présence des isoenzymes de la superfamille du cytochrome P450 (CYP). Ces réactions représentent la grande majorité des réactions de phase I ;
    - de réductions, moins fréquentes et moins étudiées ;
    - d'hydrolyse, le plus souvent par des estérases non spécifiques.

Cette première phase n'est pas obligatoire : certains médicaments peuvent directement subir la phase II.

- Les **réactions de phase II** correspondent à la conjugaison par des enzymes de type transférases, visant à rendre les métabolites encore plus hydrophiles par greffage d'un radical acétyle, sulfate, glucuronate, méthyle, glutathion, ayant une solubilité augmentée dans les fluides biologiques (urines, selles, sueur). On retrouve parmi les enzymes impliquées les N-acétlytransférases (NAT), les glutathion S-transférases (GST) ou les sulfo-transférases (SULT) ainsi que les UDP-glucuronosyltransférases (UGT).

- Les **réactions de phase III** qui permettent l'élimination de l'organisme des métabolites ainsi modifiés, par leur transport membranaire et leur expulsion hors de la cellule. Ces réactions font intervenir les transporteurs membranaires comme la famille SLC (*Solute carrier*) et la famille ABC (*ATP-binding cassette*).

2.2 Facteurs non génétiques

La variabilité interindividuelle de réponse aux médicaments peut être observée à tous les niveaux des processus pharmacodynamiques et pharmacocinétiques.

La variabilité des processus pharmacodynamiques peut être influencée par les paramètres suivants :

- la *tolérance*, qui correspond à une désensibilisation ou *down regulation* des récepteurs, qui peut être partielle (ne concerner qu'une partie des effets pharmacologiques du médicament) ou croisée (présente chez tous les médicaments d'une même classe pharmacologique).
- la *dépendance*, qui peut être d'ordre physique, avec l'apparition de troubles lors de la suppression du médicament, définissant un syndrome de sevrage et/ou psychique, caractérisée par un état compulsif poussant à consommer le médicament pour le plaisir chimique qu'il procure.
- la *chronopharmacologie*, qui concerne la modulation de l'activité pharmacologique ou toxique du médicament selon son heure d'administration. Ce phénomène s'explique par la présence des rythmes circadiens de l'organisme qui régulent un grand nombre de processus physiologiques et biologiques (sommeil, température corporelle, métabolisme hépatique, rythme cardiaque, pression artérielle, sécrétion d'acide gastrique, etc.).

Par ailleurs, différentes situations peuvent contribuer à la modification des propriétés pharmacocinétiques de la réponse aux médicaments :

- *des états physiologiques particuliers* : âge (nouveau-nés, personnes âgées), femmes enceintes et allaitantes, chez qui les paramètres cinétiques peuvent être modifiés ;
- *des situations pathologiques* : notamment la présence de comorbidités comme l'insuffisance rénale ou hépatique, qui sont les deux principales pathologies modifiant la pharmacocinétique des médicaments ; mais également l'insuffisance cardiaque, l'obésité, la

dénutrition, les grands brûlés, l'alcoolisme, les états inflammatoires et infectieux, …

- *Des interactions médicamenteuses* qui vont engendrer des phénomènes de compétition entre plusieurs médicaments au niveau du métabolisme et/ou du transport, induisant des phénomènes d'induction ou d'inhibition enzymatique.
- *des facteurs environnementaux* (tabac, alcool, alimentation, mode de vie, etc.)

Le sexe joue également un rôle dans la variabilité de réponse aux médicaments, et peut affecter aussi bien la pharmacocinétique (par exemple, *via* la modulation de la concentration plasmatique d'albumine (Wilson, 1984)) que la pharmacodynamie de médicaments (citons par exemple l'influence des récepteurs aux œstrogènes sur la réponse aux antalgiques (Qiu et al., 2008)).

Une importante variabilité des processus pharmacodynamiques et pharmacocinétiques peut conduire à une sortie de la zone thérapeutique à la posologie standard. Il est alors préconisé d'adapter le schéma posologique, en particulier pour les médicaments à marge thérapeutique étroite, afin de maintenir le patient dans une zone garantissant une efficacité pharmacologique sans risque de toxicité majeure. Cette adaptation peut être guidée par le suivi thérapeutique pharmacologique (STP), qui consiste à mesurer les concentrations plasmatiques du médicament pour ajuster la posologie de façon précise en fonction du taux obtenu ; ou par d'autres paramètres, comme l'INR (*International Normalized Ratio*) pour les anticoagulants oraux (*cf*. partie 2 de cette thèse), l'électrocardiographie (ECG) pour les antiarythmiques, etc.

2.3 Facteurs génétiques

Aux facteurs physiopathologiques et environnementaux classiques impliqués dans la variabilité de réponse aux médicaments s'ajoutent les facteurs

génétiques (Figure 1.2). Cette variabilité d'origine génétique peut être d'ordre :

- pharmacocinétique et concerner des gènes codant pour des enzymes impliquées dans le métabolisme de phase I (famille des CYP), le métabolisme de phase II (famille des UGT, GST, etc.) et le transport (famille ABC) des médicaments ;
- pharmacodynamique et concerner des gènes codant pour les éléments constitutifs ou fonctionnels des voies physiologiques ciblées par l'action des médicaments (par exemple les récepteurs des vitamines comme le gène *VDR (vitamin D (1,25- dihydroxyvitamin D3) receptor)* ou des neuromédiateurs comme le gène *DRD2 (dopamine receptor D2)*).

Ces facteurs génétiques englobent également les gènes influençant les effets thérapeutiques des médicaments sans pour autant avoir de lien direct avec la pharmacocinétique ou la pharmacodynamie des médicaments. Il s'agit par exemple des gènes codant des facteurs de transcription modulant la régulation de l'expression des gènes impliqués dans le métabolisme et le transport des médicaments, comme le gène *AHR* (*aryl-hydrocarbon receptor*) qui code pour un récepteur nucléaire[2].

Ces dernières années, de nombreux travaux ont mis en évidence le rôle majeur joué par les facteurs génétiques dans la détermination de la réponse aux médicaments. L'étude de ces facteurs a donné naissance à une sous-discipline de la pharmacologie : la pharmacogénétique.
Un résumé des facteurs intervenant dans la variabilité de la réponse aux médicaments est proposé sous forme de schéma dans la Figure 1.2.
En comparaison des différents facteurs précités, les facteurs génétiques ne varient pas au cours du temps (à part sous l'influence de mécanismes épigénétiques peu étudiés encore pour la réponse aux médicaments), ce qui

[2] Les récepteurs nucléaires appartiennent à une grande famille de facteurs de transcription activés par des ligands, le plus souvent endogènes, mais qui peuvent aussi être exogènes (xénobiotiques : polluants, médicaments). La liaison du ligand va moduler la transcription de gènes cibles selon un mécanisme aujourd'hui bien connu. Ces gènes cibles codent pour les enzymes du métabolisme et les protéines de transport membranaire qui participent aux processus de détoxification de l'organisme.

en fait de bons prédicteurs de la réponse aux médicaments et justifie l'ampleur des développements actuels autour de la pharmacogénétique.

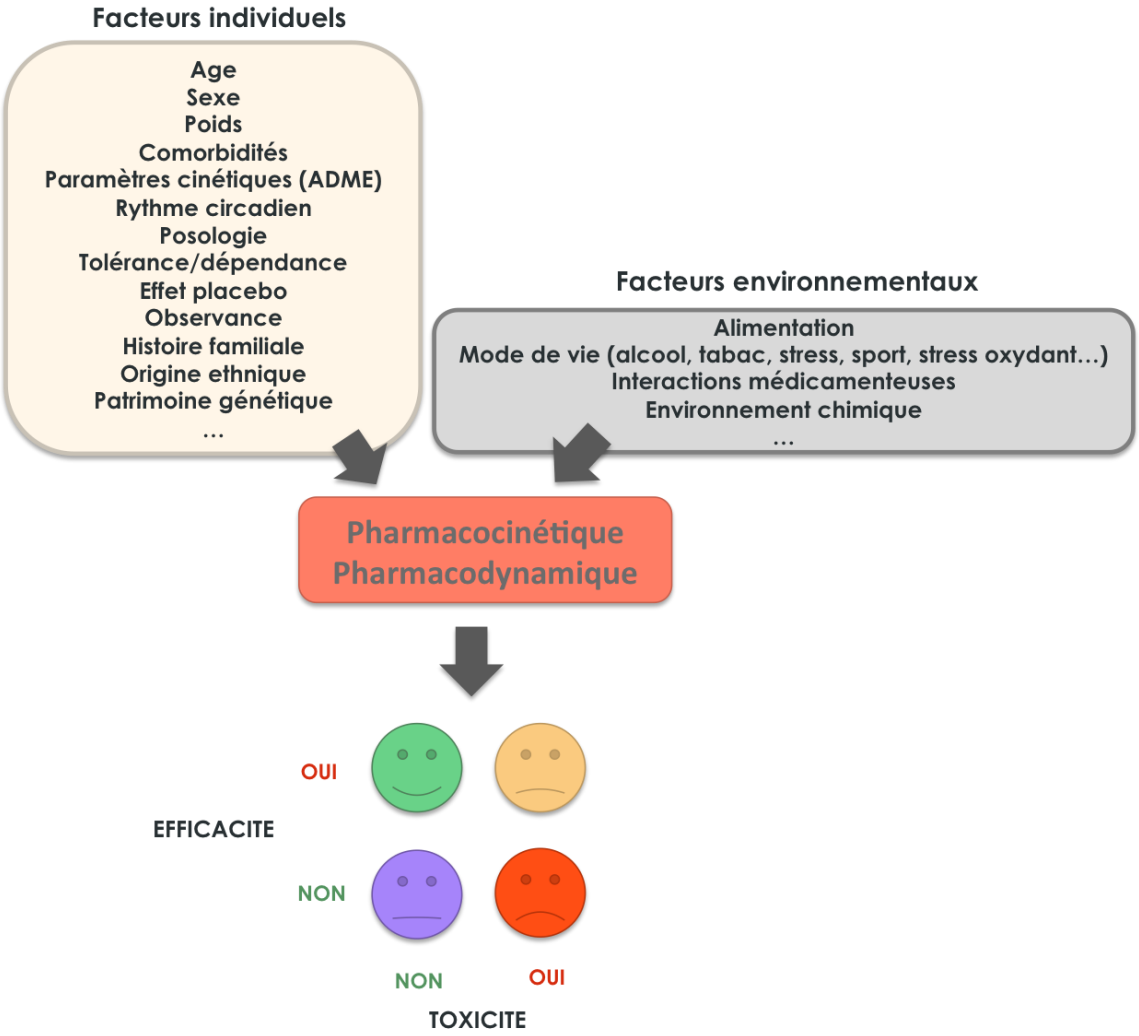


Figure 1.2 | Facteurs influençant la réponse aux médicaments.

3. La pharmacogénétique

La pharmacogénétique est l'étude de l'ensemble des mécanismes d'origine génétique intervenant dans la variabilité interindividuelle de la réponse aux médicaments. Elle a pour objectif final d'adapter les traitements médicamenteux au profil génétique de chaque patient, permettant de maximiser leur efficacité tout en minimisant leur toxicité. Le développement de la pharmacogénétique en pratique clinique courante est susceptible d'avoir un impact sans précédent sur la manière dont les maladies rares et communes seront traitées. Elle offre en effet la possibilité nouvelle de prédire et d'anticiper la réponse de l'individu au médicament et provoque de ce fait un changement de paradigme dans l'approche thérapeutique classique fondée sur l'essai-erreur.

La pharmacogénétique vise à améliorer l'efficacité et la sûreté des traitements médicamenteux en apportant des éléments d'aide à trois grands niveaux de la stratégie thérapeutique (Salari et al., 2012) :

(1) En permettant de stratifier les patients selon leur profil génétique, il est possible dans certaines maladies, notamment des cancers, d'orienter la stratégie thérapeutique en administrant un médicament spécifiquement adapté à la maladie du patient, qui est donc susceptible d'avoir un bon niveau de réponse. Ainsi, on dispose, en particulier en oncologie, de thérapies ciblées dont l'utilisation est restreinte à des sujets porteurs d'altérations génétiques spécifiques, comme c'est le cas par exemple des leucémies myéloïdes chroniques exprimant le gène de fusion *BCR/ABL* produisant une protéine à activité tyrosine kinase inhibée spécifiquement par l'imatinib (Glivec®).

(2) En prédisant l'efficacité du médicament selon le génotype du patient pour des polymorphismes génétiques modifiant les propriétés pharmacodynamiques ou pharmacocinétiques du médicament qui peuvent entraîner une sortie de la zone thérapeutique à la posologie standard.

(3) En anticipant et évitant la survenue d'effets indésirables chez certains sujets porteurs de génotypes conférant un risque de toxicité augmenté pour certains médicaments.

Dans ces deux derniers cas, la connaissance du profil génétique du patient peut conduire à une surveillance biologique et clinique particulière ou à une stratégie thérapeutique modifiée (adaptation posologique, choix de la molécule).

La perspective de l'individualisation des traitements médicamenteux passe par le développement de tests simples utilisables en pratique médicale courante, incorporant l'information génétique du patient. Ces tests pharmacogénétiques font appel au développement de biomarqueurs compagnons, essentiels au suivi de la réponse pharmacologique des médicaments dans le contexte de la médecine personnalisée (Encadré 1.3) (Biomarkers Definitions Working Group., 2001; Sim and Ingelman-Sundberg, 2011).

Encadré 1.3 I Notion de biomarqueurs en pharmacogénétique

De façon générale, un biomarqueur peut être défini comme un outil de mesure (biochimique ou moléculaire), objectif, précis et reproductible, des caractéristiques d'un processus biologique, physiologique ou pathologique ou de la réponse pharmacologique à un traitement.
Les biomarqueurs font partie des pratiques médicales depuis de nombreuses années, mais leur développement, fortement lié à celui des technologies, s'est accéléré dans la dernière décennie, suscitant de grands espoirs dans l'implémentation de la médecine personnalisée.

Leur intérêt en pharmacogénétique réside à différents niveaux :

- Ils sont utilisés en association avec les thérapies médicamenteuses (« biomarqueurs compagnons ») pour améliorer leur qualité et efficacité, et diminuer leur toxicité ; en permettant de :
    - dépister les patients chez qui le traitement va être le plus efficace (par exemple : marqueur diagnostique d'une maladie pour laquelle on dispose de thérapies ciblées) ou au contraire les non répondeurs ;
    - faciliter le suivi du traitement médicamenteux en guidant le dosage pour rester dans la zone thérapeutique et limiter la variation interindividuelle de la réponse pharmacologique au médicament.

- L'association médicament/biomarqueur a des intérêts également d'ordre économique en :
    - transformant le développement et la mise à disposition d'un nouveau médicament en un processus plus rapide (élimination des molécules candidates les moins efficaces en phase précoce et anticipation des échecs en phase avancée du développement) ; et moins coûteux (réduction du nombre d'échecs thérapeutiques) ;
    - permettant de récupérer les médicaments retirés du marché du fait de leur toxicité.

- Par ailleurs, ils contribuent à améliorer les connaissances de pharmacologie clinique et fournissent une aide pour la conception des essais cliniques, qui évalueront définitivement la sécurité et l'efficacité de la molécule.

Les biomarqueurs génétiques permettant de prédire la réponse aux médicaments ou leur toxicité sont surtout des variants génétiques situés dans les gènes impliqués dans le métabolisme et le transport des médicaments, les gènes codant pour les cibles thérapeutiques, ou les gènes du système HLA.
A l'heure actuelle, l'oncologie demeure le domaine où les associations biomarqueurs/médicaments sont les plus développées.

3.1 Histoire de la pharmacogénétique

La toute première description d'une variation interindividuelle de réponse à un xénobiotique, qui s'avérera plus tard être liée à une anomalie de réponse à un médicament, remonte au cinquième siècle avant J.C. : en observant chez certains individus d'origine méditerranéenne des crises d'hémolyses aigües suite à l'ingestion de fèves (pathologie appelée de ce fait « favisme »), le philosophe et mathématicien grec Pythagore avait déconseillé cet aliment en raison de son effet pathologique potentiel (Cappellini and Fiorelli, 2008). Il déclara : « *ce qui peut être nourriture pour certains, peut être poison violent pour d'autres* ». Ce n'est que beaucoup plus tard que la cause génétique de cette variabilité métabolique fut identifiée, suite à l'observation d'anémies hémolytiques similaires aux anémies faviques après l'administration d'un médicament (primaquine). A Londres, le médecin anglais Sir Archibald E. Garrod (1857-1936), s'intéressant aux pathologies liées à des déséquilibres biochimiques, fait l'hypothèse que certaines maladies, comme l'alcaptonurie[3], sont en réalité des troubles héréditaires du métabolisme. En 1902, il postule l'existence d'une mutation dans un gène codant pour cette enzyme qui serait à l'origine de l'alcaptonurie et développe le concept « d'individualité chimique » (Garrod, 2002). Il publie en 1909 son livre *Inborn Errors of Metabolism*, dans lequel il propose que des facteurs génétiques pourraient contribuer aux variations interindividuelles dans le métabolisme et la réponse aux médicaments.

Première expérience de pharmacogénétique

La première expérience relevant du domaine de la pharmacogénétique n'a pas concerné une variation interindividuelle dans la réponse à un médicament, mais une capacité à déceler un composé chimique étranger (xénobiotique) : en 1931, le chimiste américain Arthur Fox découvre que la molécule sur laquelle il travaille, le phénylthiocarbamide (PTC[4]) a un goût très amer pour certaines personnes, alors qu'elle n'en a aucun pour d'autres, notamment pour lui-même (Fox, 1932). La même année, l'américain Snyder démontre que cette particularité se transmet selon un mode

[3] Maladie provoquée par un déficit en homogentysate dioxygénase, enzyme impliquée dans le métabolisme de l'acide homogentisique, métabolite dérivé de la tyrosine

[4] Composé organique amer fabriqué par certaines plantes qui l'utilisent comme répulsif pour se protéger des herbivores.

mendélien récessif (Snyder, 1931). Les « non-goûteurs » ne possèdent pas le récepteur au PTC, codé par le gène *TAS2R38* (Kim et al., 2003). Aujourd'hui, le test déterminant la sensibilité au PTC constitue l'un des tests génétiques les plus communs chez l'homme.

Premiers exemples pharmacogénétique de variabilité interindividuelle dans la réponse aux médicaments

Dans les années 1950, trois exemples ont démontré clairement l'existence de variations génétiquement déterminées dans l'activité d'enzymes étant à l'origine d'effets indésirables aux médicaments, comme Garrod l'avait prédit.

Le premier concerne la primaquine, un antipaludique dont la consommation déclenchait des crises hémolytiques aiguës chez certains soldats d'origine africaine durant la Seconde Guerre Mondiale (CLAYMAN et al., 1952). Il a été démontré plus tard que cette réaction est due à un déficit en glucose-6-phosphate déshydrogénase (G6PD), qui entraîne une destruction des globules rouges (ALVING et al., 1956). C'est cette même anomalie génétique qui est à l'origine du favisme décrit plusieurs siècles avant par Pythagore.

Le second exemple concerne le déficit en pseudocholinestérases plasmatiques, découvert par Kalow en 1957. Ce déficit rare entraîne des réactions de paralysie prolongée après injection de succinylcholine (ou suxaméthonium), un curare dépolarisant utilisé en médecine anesthésique (KALOW and STARON, 1957). Il se transmet selon un mode autosomique récessif par mutation du gène *BCHE* (McGuire et al., 1989).

Le dernier exemple, sans doute le plus connu, concerne l'isoniazide, molécule introduite en 1952 pour le traitement de la tuberculose, pour laquelle a été observé une incidence élevée de neuropathies périphériques comme réaction indésirable, du fait d'une accumulation dans l'organisme de certains patients de la forme non transformée de la molécule (Hughes et al., 1954). Ces patients, éliminant de façon moins efficace le médicament, sont alors appelés « inactivateurs lents » de l'isoniazide ; ils constituent une proportion non négligeable (environ 40 %) des patients traités avec ce médicament. Quelques années plus tard, Evans et White montrent que la plus faible vitesse d'élimination de ce médicament a pour origine une réduction de la vitesse de *N*-acétylation de la molécule active dans le foie

(Evans et White, 1964). Les inactivateurs deviennent alors les *acétyleurs* lents de l'isoniazide. Ce n'est que trente ans plus tard que le gène *NAT2* est identifié comme le site du polymorphisme d'acétylation chez l'homme (Blum et al., 1991).

Avènement de la pharmacogénétique en tant que discipline distincte

Reconnaissant l'importance de ces découvertes, Arno Motulsky les relate pour la première fois en 1957 dans un article intitulé *Drug reactions enzymes, and biochemical genetics,* posant ainsi les fondations intellectuelles de la pharmacogénétique, qui devient alors une discipline à part entière (MOTULSKY, 1957). Le terme pharmacogénétique est inventé par l'allemand Friedrich Vogel en 1959 (Vogel, 1959). Werner Kalow publie en 1962 la première monographie exhaustive des couples médicaments/gènes connus en pharmacogénétique (Kalow, 1962). La première conférence internationale de pharmacogénétique se déroule en 1967 à l'Académie des sciences de New York. Les études de pharmacogénétique réalisées sur des jumeaux à partir de la fin des années 60 confirment l'hypothèse apportée par Garrod que les facteurs génétiques jouent un rôle majeur dans la variabilité de réponse aux médicaments (Vesell, 1978).

Depuis, le champ de la pharmacogénétique s'est considérablement développé, comme l'atteste le nombre croissant, voire exponentiel, de publications qui y sont consacrées depuis une vingtaine d'années (Figure 1.3). Le premier journal consacré à la pharmacogénétique est créé en 1991 (*Pharmacogenetics*). Aujourd'hui, de très nombreux exemples de polymorphismes génétiques à l'origine de variations interindividuelles dans la réponse aux médicaments ont été rapportés. On compte à l'heure actuelle, plus de 2000 gènes annotés dans les bases de données de pharmacogénétique (Salari et al., 2012). Parmi les bases existantes, *PharmGKB* (pour *Pharmacogenomics Knowledge Base*) est la plus connue et la plus complète (Klein et al., 2001). Cependant il en existe de nombreuses autres, dont : *CYP-allele database, NAT-allele database, PMT database, TP-search database, UGT-allele database, PharmaADME* (Sim et al., 2011),

PharmGED (Zheng et al., 2007), e-PKgene (Hachad et al., 2011), ou encore PACdb (Gamazon et al., 2010).

De plus, le développement récent des technologies de génomique de nouvelle génération marque l'entrée dans l'ère de la pharmacogénomique, qui diffère de la pharmacogénétique en terme d'échelle (Encadré 1.4).

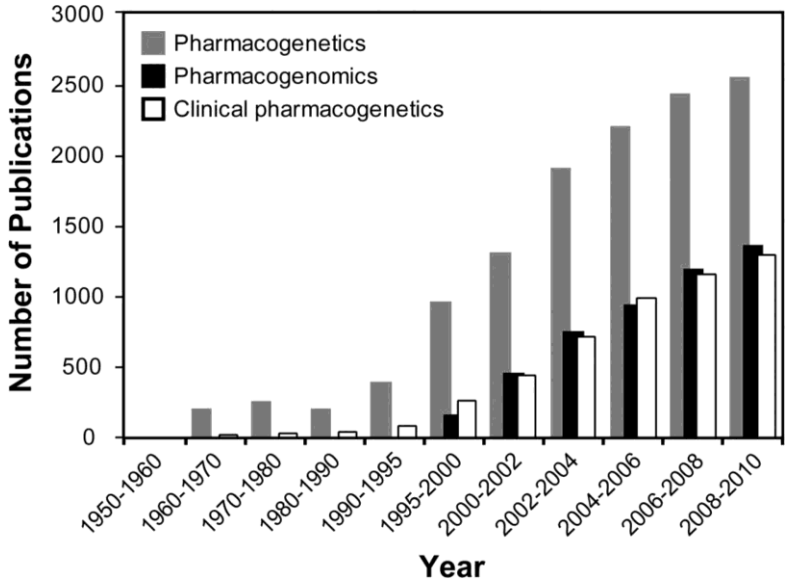


Figure 1.3 | Nombre de publications contenant les termes « *pharmacogenetics* », « *pharmacogenomics* » et « *clinical pharmacogenetics* », tiré de (Scott, 2011).

Encadré 1.4 I Distinction des termes pharmacogénétique / pharmacogénomique

Depuis l'introduction du mot *pharmacogénomique* en 1997 (Marshall, 1997), ces deux termes ont tendance à être employés de façon interchangeable, et aucune définition précise ne fait consensus à l'heure actuelle.

- Généralement, la pharmacogénétique est considérée comme l'étude de l'influence du statut génétique sur la réponse aux médicaments. On la distingue de la pharmacogénomique qui désigne non pas l'étude des modifications de la séquence génétique, mais d'un point de vue plus vaste celle du profil d'expression des gènes impliqués dans la susceptibilité aux phénotypes de la réponse aux médicaments aux niveaux cellulaire, tissulaire, individuel, ou populationnel. Elle recouvre donc des niveaux d'analyse supérieurs à celui de l'ADN, en s'intéressant également à l'ARN et aux protéines. La pharmacogénomique recouvre alors le champ d'application des technologies de la génomique de nouvelle génération.

- Certains associent à ces deux termes une différence d'échelle uniquement : la pharmacogénétique se réfère à l'étude d'un seul gène candidat tandis que la pharmacogénomique implique l'étude de nombreux gènes, voire du génome entier et de réseaux de gènes.

- D'autres suggèrent que la pharmacogénomique concerne davantage le développement de nouveaux médicaments et la caractérisation des médicaments existants que la pharmacogénétique, plus centrée sur la compréhension de la variabilité de la réponse aux médicaments.

- En cancérologie, la pharmacogénétique se réfère historiquement à des mutations germinales et la pharmacogénomique à des mutations somatiques dans l'ADN tumoral, conduisant à une modification de la réponse aux médicaments de chimiothérapie.

Diffusion des connaissances en pharmacogénétique

Différents efforts ont été entrepris pour favoriser une organisation et une diffusion des connaissances en pharmacogénétique, l'objectif étant de faciliter leur intégration aux données cliniques et biologiques traditionnelles afin d'optimiser les traitements en pratique médicale courante. Ainsi, en 2000 a été créé aux États-Unis, avec le soutien du *National Institute of Health* (NIH),

le *Pharmacogenomics Research Network* (PGRN), réseau national collaboratif de groupes scientifiques visant à mieux comprendre la façon dont le patrimoine génétique d'un individu influence sa réponse aux médicaments. Une base de données publique PharmGKB (http://www.pharmgkb.org) a également été mise en place qui collecte, annote, agrège, organise, et résume l'ensemble des connaissances disponibles concernant l'impact des variations génétiques sur la réponse aux médicaments, à destination des cliniciens et des chercheurs. La collaboration entre le PGRN et PharmGKB a conduit à la formation en 2009 du *Clinical Pharmacogenetics Implementation Consortium* (CPIC), dont l'objectif majeur est de favoriser l'implémentation des tests pharmacogénétiques en routine clinique. Il émet des directives (*guidelines*) basées sur l'information génotypique, publiées dans le journal *Clinical Pharmacology and Therapeutics* et diffusées sur le site internet de PharmGKB. Ces directives sont conçues de telle sorte qu'elles contiennent, à travers un format standardisé, les informations nécessaires pour aider les cliniciens à mieux comprendre l'intérêt d'utiliser des tests pharmacogénétiques en vue d'optimiser les traitements médicamenteux et pour leur apporter une aide dans l'interprétation des résultats de ces tests.

FDA labelling

La FDA (*Food and Drug Administration*) a été la première autorité compétente à prendre position dès 2003, en mentionnant explicitement dans les résumés des caractéristiques du produit (RCP) la nécessité de réaliser des tests pharmacogénétiques avant l'utilisation de certains médicaments considérés comme « à risque » d'entraîner des effets indésirables graves voire mortels chez certains patients. En 2005, 40 % des médicaments approuvés par la FDA contiennent des informations pharmacogénétiques dans leur RCP (Figure 1.4).

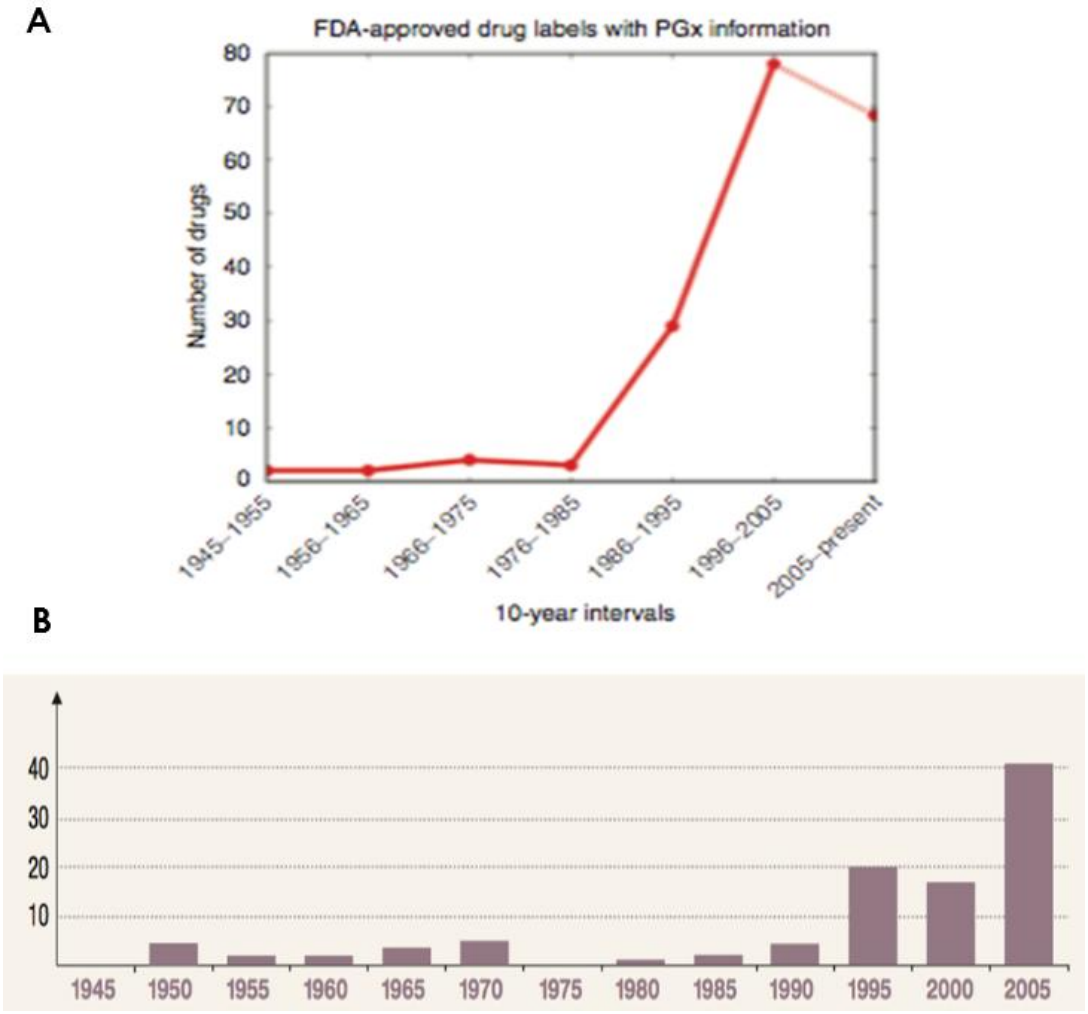


Figure 1.4 | (A) Nombre de médicaments possédant un label pharmacogénétique approuvé par la FDA par décennie (tiré de Cordero rev 2012). (B) Part des médicaments approuvés contenant des informations pharmacogénétiques (en % du nombre total). ***Source : Analyse Bionest Partners, FDA, Rapport de la commission EU sur la pharmacogénétique et pharmacogénomique, 2006.***

Gènes d'intérêt majeur en pharmacogénétique

La base de données PharmGKB distingue parmi les milliers de gènes découverts par les études de pharmacogénétique une cinquantaine de gènes présentant un intérêt majeur pour expliquer les différences de réponse aux médicaments entre les individus : les *Very Important Pharmacogenes* (VIP) (Figure 1.5).

Ces gènes se répartissent en quatre grandes catégories : (1) les gènes codant pour les enzymes du métabolisme des médicaments ; (2) ceux

codant pour les transporteurs membranaires des médicaments ; (3) ceux codant pour les récepteurs ou sites « cibles » des médicaments et enfin (4) les gènes modificateurs qui affectent l'expression des gènes ou protéines impliqués dans les trois processus cités ci-dessus.

Pour chacun de ces gènes VIP, la base de données PharmGKB fournit un résumé (le *VIP summary*) dans lequel sont récapitulées l'ensemble des connaissances pharmacogénétiques actuelles relatives à ce gène, incluant le détail des variations génétiques ayant été associées aux phénotypes de réponse aux médicaments.

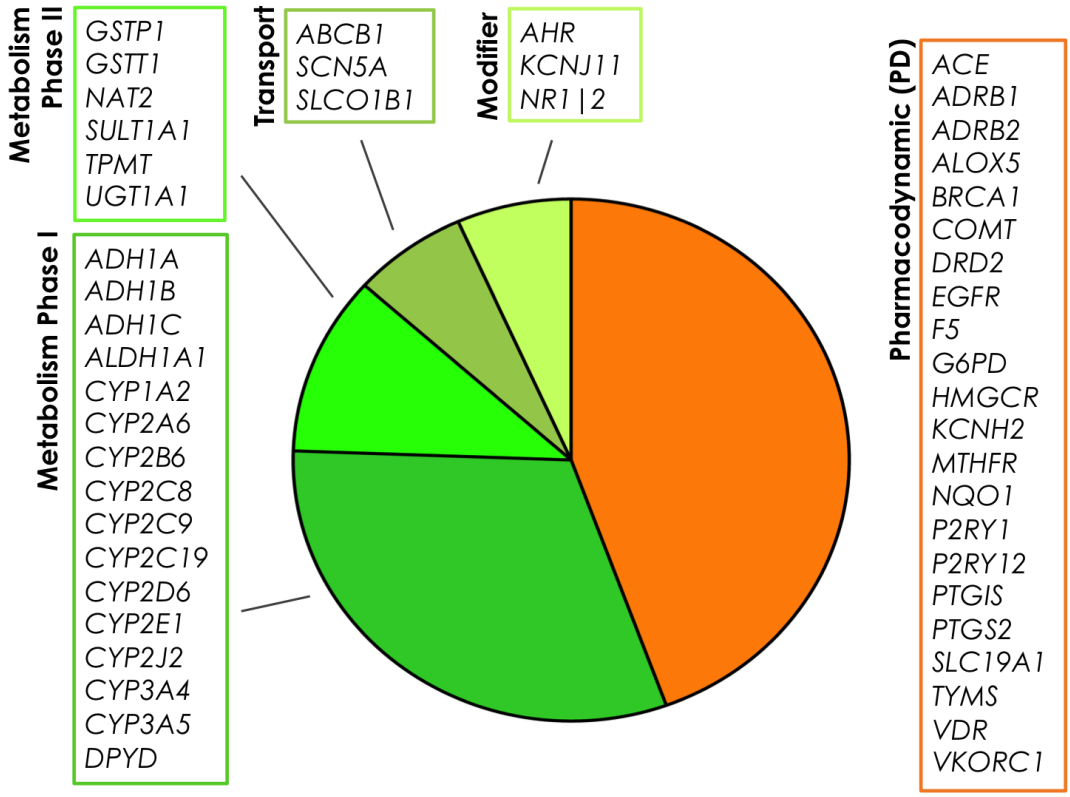


Figure 1.5 | Liste des 50 gènes VIP (*Very Important Pharmacogenes*) annotés dans la base de données PharmGKB (www.pharmgkb.org) en mars 2014.

Possibilité d'implémentation rapide en clinique des connaissances pharmacogénétiques

L'identification des variants génétiques jouant un rôle majeur dans la réponse aux médicaments est plus susceptible de conduire à des applications cliniques concrètes et immédiates que les variants impliqués dans la susceptibilité aux maladies communes (Goldstein et al., 2003). En effet, en comparaison aux maladies communes, il semble que certains phénotypes de la réponse aux médicaments soient déterminés par une composante génétique et physiologique plus simple. En conséquence, les associations avec les génotypes sous-jacents peuvent avoir une utilité diagnostique directe en clinique. Par exemple les thérapies médicamenteuses déjà existantes peuvent être améliorées grâce à l'usage de biomarqueurs génétiques permettant d'éviter la survenue des effets indésirables sévères ou de prédire l'efficacité thérapeutique. A l'inverse, les variants communs de prédisposition aux maladies complexes ont le plus souvent une valeur prédictive faible, et si leur identification peut permettre d'indiquer une nouvelle cible thérapeutique potentielle, il faut un certain temps pour développer de nouvelles molécules thérapeutiques capables d'agir sur ces cibles.

3.2 Méthodes de la pharmacogénétique

Les phénotypes de la réponse aux médicaments sont le plus souvent des traits quantitatifs, déterminés par de multiples facteurs génétiques et environnementaux. Les mêmes méthodes que celles utilisées en génétique pour élucider la base génétique des traits multifactoriels sont donc employées pour identifier les variants d'intérêt en pharmacogénétique. En revanche, il est à noter que les analyses d'association génétique conduites sur des populations d'individus non apparentés (constituées par exemple de répondeurs et de non répondeurs à un traitement médicamenteux) sont beaucoup plus couramment utilisées, compte tenu de la difficulté de disposer de phénotypes liés à la réponse aux médicaments chez les différents membres d'une famille. Cela est surtout vrai pour certains médicaments,

comme les molécules utilisées en chimiothérapie anticancéreuse, qui sont rarement administrés à plusieurs membres d'une même famille.

Dans le cadre de ces études d'association génétique, et à l'instar des études visant à identifier les gènes impliqués dans les maladies complexes, deux grands types d'approches sont employés pour identifier les variants d'intérêt en pharmacogénétique :

- L'*approche gène candidat*, qui consiste à rechercher et analyser les variants génétiques dans des gènes dont la fonction pourrait jouer un rôle dans la détermination du trait étudié ;
- La stratégie *génome entier* qui est une approche « agnostique » sans hypothèse préalable sur les gènes d'intérêt, permettant de découvrir la fonction de nouveaux gènes.

Études gènes candidats

Les facteurs génétiques impliqués dans les phénotypes de la réponse aux médicaments ont d'abord été recherchés par des études gènes candidats. Ces études ont globalement conduit à de meilleurs résultats en pharmacogénétique que dans le cas des maladies complexes, probablement parce que l'on disposait le plus souvent de meilleurs candidats. En effet, les principales enzymes impliquées dans le métabolisme d'un médicament et les principales cibles pharmacologiques de ce dernier sont bien souvent connues, permettant de disposer de gènes candidats évidents qui se sont souvent avérés être effectivement des déterminants importants de la réponse au médicament étudié. Ces études ont ainsi permis d'identifier de nombreux gènes d'intérêt majeur en pharmacogénétique, conduisant parfois au développement de tests pharmacogénétiques d'utilisation courante en pratique clinique. Citons par exemple le cas des gènes *TPMT*, *UGT1A1* et *VKORC1*.

Le portage d'un ou plusieurs des allèles *TPMT*2* et **3* entraîne une perte de fonction du gène *TPMT* et un déficit enzymatique en thiopurine S-méthyltransférase, augmentant alors le risque d'aplasie médullaire dans les traitements par les thiopurines telles que l'azathioprine, la 6-mercaptopurine, ou la 6-thioguanine. Ces molécules antimétaboliques sont utilisées pour leurs propriétés immunosuppressives dans les transplantations d'organes et les

maladies auto-immunes et pour leurs propriétés cytotoxiques dans certaines leucémies. Ce biomarqueur a fait l'objet de la toute première directive du CPIC, recommandant aux médecins l'usage du test de génotypage et les aidant à interpréter son résultat (Relling et al., 2011, 2013).
Les sujets porteurs à l'état homozygote de l'allèle *UGT1A1*28* du gène *UGT1A1* codant pour l'*uridine diphosphate glucuronosyltransférase 1A,* sont à risque accru de neutropénie lors d'un traitement par l'irinotécan, agent antinéoplasique utilisé dans la prise en charge des cancers colorectaux (Innocenti et al., 2009).
Les anticoagulants oraux de type antivitamine K (AVK) disposent dorénavant de tests pharmacogénétiques intégrant le statut génétique des deux gènes *CYP2C9* et *VKORC1*, permettant d'adapter les posologies initiales d'AVK à administrer au patient. Les algorithmes pharmacogénétiques de calcul de la dose testent la présence des allèles *CYP2C9*2* et *CYP2C9*3* (associés à des niveaux de métabolisme abaissés du CYP2C9) et du variant fonctionnel rs9923231 du gène *VKORC1* (qui confère une sensibilité augmentée aux AVK) (*cf.* partie 2) (Johnson et al., 2011).

Bien que les études gène candidat aient permis d'améliorer notre compréhension de la base génétique de la variabilité de réponse aux médicaments, elles se sont essentiellement focalisées sur les gènes impliqués dans le métabolisme des médicaments ou ceux codant pour les cibles pharmacologiques (récepteurs ou enzymes), ne permettant d'explorer qu'une partie de cette composante génétique. Les phénomènes de survenue d'effets indésirables sont particulièrement concernés puisqu'ils font intervenir des gènes impliqués dans un grand nombre de réactions physiologiques, impliquant notamment les fonctions immunitaires et mitochondriales (Daly, 2010). Citons par exemple l'implication des gènes *HLA* dans les réactions d'hypersensibilité à certains médicaments, comme l'allèle *HLA-B*5701* avec l'abacavir (Mallal et al., 2002).

Études d'association pangénomiques (GWAS)

Une alternative à ce problème est apportée par les études d'association pangénomiques (*Genome Wide Association Studies* GWAS) qui permettent

d'analyser l'ensemble du génome sans poser d'hypothèse *a priori* sur les gènes en cause dans le phénotype et d'explorer l'effet de plusieurs gènes sur un même phénotype.

Cette approche s'est révélée être puissante, car elle a permis de détecter de nombreuses associations significatives à l'échelle du génome entier, qui sont aujourd'hui utiles en clinique. Citons par exemple, parmi les GWAS étudiant les phénotypes de toxicité médicamenteuse, la découverte de l'association du gène de transport hépatique *SLCO1B1* (*solute carrier organic anion transporter family, member 1B1*) avec un risque augmenté de myopathies induites par les statines, molécules fréquemment utilisées dans le traitement de l'hypercholestérolémie (SEARCH Collaborative Group et al., 2008). Un variant non synonyme rs4149056 (définissant l'allèle *SLCO1B1*5*) cause une réduction de la fonction de transport hépatique de ces molécules (Iwai et al., 2004). Les individus homozygotes pour ce variant sont plus à risque de développer des atteintes musculaires graves voire mortelles. Ce résultat a conduit à une modification du label de la FDA et une directive du CPIC recommandant de tester la présence de ce variant génétique avant de démarrer une thérapie par simvastatine, afin d'adapter le dosage du médicament et le suivi clinique et biologique de la fonction musculaire du patient (Wilke et al., 2012). Un autre exemple est la découverte de l'association du variant non synonyme rs4244285, définissant l'allèle *CYP2C19*2*, dans la survenue d'évènements cardiovasculaires indésirables sévères lors d'un traitement par clopidogrel, un antiplaquettaire (Mega et al., 2010; Shuldiner et al., 2009). Cette association a donné lieu à une directive du CPIC recommandant l'usage du test de génotypage, pour guider le choix de la molécule à administrer au patient (Scott et al., 2011).

De manière intéressante, certaines études ont détecté des variants pharmacogénétiques ayant des effets particulièrement forts (odds ratios élevés), résultat qui contraste avec ceux des GWAS consacrées aux maladies complexes (Meyer et al., 2013). Cette différence marquée suggère qu'il existe des différences entre les déterminants génétiques des traits pharmacogénétiques et ceux des maladies complexes, les variants pharmacogénétiques étant associés aux phénotypes d'intérêt avec des

odds ratios globalement plus élevés, ce qui facilite leur identification (Meyer et al., 2013). Toutefois cette observation n'est pas généralisable à l'ensemble des études GWAS consacrées aux phénotypes de la réponse aux médicaments, qui n'ont parfois rien détecté, ce qui s'explique notamment par le fait qu'elles font face à des problèmes différents de ceux des GWAS consacrées aux maladies complexes (Encadré 1.5).

Encadré 1.5 I Difficultés rencontrées dans les GWAS en pharmacogénétique

Les GWAS en pharmacogénétique rencontrent des difficultés spécifiques liées à :

- La difficulté de constituer des échantillons de grande taille : les patients non répondeurs à un médicament étant moins fréquents que les répondeurs, et l'incidence des effets indésirables faibles.
- La complexité de la caractérisation précise du phénotype, qui requiert une définition clinique claire de la réponse aux médicaments.
- Les GWAS en pharmacogénétique sont souvent réalisées lors d'essais cliniques, longs et coûteux à mettre en place. Il est alors plus difficile de procéder à des études de réplication, du fait des obstacles financiers et parfois éthiques pour la constitution de l'échantillon.
- La puissance de détection des variants génétiques est restreinte aux variants ayant un effet fort du fait des tailles d'échantillons limitées. Il est probable que des effets plus modérés voire faibles, comme ceux observés dans les maladies complexes, seraient détectés avec de plus grands échantillons.

La puissance des études GWAS est néanmoins limitée par les puces de génotypage employées, qui ne fournissent pas une couverture du génome homogène et suffisante pour capturer toute la variabilité du génome, en particulier pour les populations africaines, caractérisées par une plus grande diversité génétique et de plus faibles niveaux de déséquilibre de liaison (Teo et al., 2010). En conséquence, si les GWAS représentent une bonne approche pour localiser les gènes d'intérêt, l'identification des variants causaux doit passer dans un second temps par des études de séquençage. Ce problème est particulièrement vrai pour les gènes de la réponse aux médicaments dont la variation est mal représentée dans les puces utilisées dans les GWAS de

pharmacogénétique, y compris les plus denses (Gamazon et al., 2012). Différentes raisons sont à l'origine de ce constat :

- D'une part parce qu'elles ne permettent pas d'avoir accès aux variants rares et peu fréquents qui, bien que mal explorés encore, sont susceptibles d'avoir des effets importants sur les phénotypes de la pharmacogénétique (Cordero and Ashley, 2012). En effet une étude récente a montré que des variants rares dans le gène *SLCO1B1* ont des effets plus importants sur la clairance du méthotrexate que des variants communs (Ramsey et al., 2012). Il a également été montré que les variants rares ayant un fort potentiel délétère sont très nombreux dans les gènes codant pour les récepteurs des médicaments (Nelson et al., 2012).
- D'autre part parce que les gènes de la réponse aux médicaments sont parfois localisés dans des régions du génome chevauchées par la présence de variants structuraux (comme les CNVs (*Copy Number Variations*) et les insertions, délétions, etc.) qui, en l'absence de données de séquençage, limitent la possibilité de détection de la variabilité de ces gènes (Gamazon et al., 2011). C'est par exemple le cas du gène *SULT1A1* qui possède un CNV important sur le plan fonctionnel (Hebbring et al., 2008), du gène *CYP2D6* et des gènes de la famille GST (*GSTT1* et *GSTM1*) qui en possèdent plusieurs (Johansson and Ingelman-Sundberg, 2008).
- Enfin parce que certains gènes sont situées dans des régions présentant une grande homologie avec d'autres régions du génome, comme les gènes de la superfamille des cytochromes P450, ce qui complique leur étude par les méthodes GWAS.

L'avènement des données de séquençage génome entier peut permettre de pallier aux difficultés rencontrées par les études utilisant des données de génotypage. Elles présentent l'avantage, nous le verrons dans le chapitre 2, de donner accès à l'exhaustivité de la variabilité du génome humain (incluant les variants rares et peu fréquents), sans biais de découverte des variants génétiques dans certaines populations du monde.

4. Variabilité inter-populationnelle de la réponse aux médicaments

Werner Kalow a été le premier, dans son article *Ethnic differences in drug metabolism* publié en 1982, à attirer l'attention sur l'existence d'une hétérogénéité de réponse entre les populations humaines aux traitements médicamenteux (Kalow, 1982). Depuis, de multiples exemples de différences de réponse aux médicaments entre individus de différentes origines ethniques ont été rapportés dans la littérature, qu'il s'agisse de différences dans l'efficacité des médicaments ou dans l'incidence de leurs effets indésirables. Il est désormais admis que les facteurs populationnels jouent un rôle important dans la variabilité de réponse aux médicaments.

4.1 Exemples de médicaments dont la réponse présente une variabilité inter-populationnelle

Un grand nombre de médicaments couramment utilisés en pratique médicale pour le traitement de nombreuses maladies communes présentent une variabilité inter-populationnelle dans la réponse thérapeutique, qui peut se traduire en termes de :

- toxicité différentielle : c'est le cas par exemple pour la carbamazépine, utilisée dans le traitement de l'épilepsie, qui est le principal médicament inducteur des réactions cutanées sévères dont le syndrome de Stevens-Johnson (Chung et al., 2010). L'incidence de ces effets indésirables graves dus à ce médicament varie entre les populations humaines : ils sont notamment trois fois plus fréquents en Asie qu'en Europe (Chung et al., 2010). Cette différence s'explique en partie par la différence de fréquence de l'allèle *HLA-B*1502* entre les populations humaines, fortement associé à ces réactions toxiques (Chung et al., 2004). Nous pouvons également citer la variabilité inter-populationnelle dans le taux de survenue d'effets indésirables liés à la toxicité de certains agents anticancéreux (tels que l'anthracycline, le cisplatine et le fluorouracile) plus élevés dans les populations africaines que dans les populations européennes (Aminkeng et al., 2014). De plus, des taux de toxicité plus

faibles en réponse à différentes molécules (fluorouracile, leucovorine, oxaliplatine et irinotécan) utilisées dans le traitement du cancer colorectal ont été observés chez des individus d'origine asiatique en comparaison à des individus d'origine européenne (Loh et al., 2013).

- d'efficacité différentielle : c'est le cas par exemple du tacrolimus, indiqué dans la prévention du rejet de greffe en transplantation d'organes, qui requiert chez les patients Africains-Américains des doses plus élevées que les patients d'origine européenne et hispanique pour obtenir la même efficacité (Mancinelli et al., 2001). Cette différence s'explique par une biodisponibilité orale réduite de ce médicament chez ses patients, entraînant une réduction des concentrations plasmatiques (Mancinelli et al., 2001). Cette variabilité pharmacocinétique est sans doute reliée à des différences dans l'expression de la glycoprotéine de transport P-gp et du cytochrome CYP3A4 qui prennent en charge ce médicament dans l'organisme (Yasuda et al., 2008). Des variants génétiques situés dans les deux gènes codant pour ces protéines (respectivement *ABCB1* et *CYP3A4*) présentent une distribution hétérogène entre les populations humaines (Kim et al., 2012; Li et al., 2011), comme nous le verrons dans la partie 3. Un autre exemple est la warfarine, un AVK couramment utilisé dans le traitement des pathologies thrombotiques, pour laquelle des différences pharmacodynamiques sont à l'origine d'une importante variabilité de réponse entre les populations humaines. En effet, des doses moyennes quotidiennes de warfarine de 6 mg pour les individus Africains-Américains, 5 mg pour les Européens, et 3,5 mg pour les Asiatiques sont requises pour obtenir le même degré d'anticoagulation (Schelleman et al., 2008). Par ailleurs, une variabilité de réponse aux bêta-bloquants utilisés dans le traitement de l'hypertension artérielle, dont le propranolol (Venter and Joubert, 1984), le métoprolol (Rutledge et al., 1989) et l'aténolol (Materson et al., 1993), a été observée chez des patients d'origine européenne et africaine. Étant donné que cette variabilité de sensibilité à ces médicaments se manifeste à concentrations plasmatiques égales, il est possible qu'elle soit causée par des facteurs pharmacodynamiques, tel qu'une variation de la densité en récepteurs bêta-adrénergiques (Wood, 1998).

Entre 2005 et 2007, 50 % des nouvelles molécules approuvées par la FDA ont bénéficié d'un label pharmacogénétique intégrant l'information sur l'origine ethnique du patient (Yasuda et al., 2008).

Le cas du BiDil®

En juin 2005, la FDA, délivrait une surprenante autorisation de mise sur le marché d'un premier médicament « ethnique », le BiDil®, qui déclencha un tollé général dans la communauté scientifique et médicale. Composé de deux molécules couramment utilisées dans le traitement de l'insuffisance cardiaque (chlorhydrate d'hydralazine et dinitrate d'isosorbide), ce médicament avait d'abord été jugé inefficace sur un échantillon de patients choisis dans l'ensemble de la population américaine (Cohn et al., 1991). Un nouvel essai clinique, effectué cette fois exclusivement sur un échantillon de patients Africains-Américains révélait, en revanche, une certaine efficacité, réduisant la mortalité de 43 % et l'hospitalisation de 39 % quand il était pris en combinaison avec une thérapie standard (Taylor et al., 2004). Au vu des résultats de cette étude, la FDA, délivrait son autorisation, pour le traitement de ce segment particulier de la population américaine, les noirs originaires d'Afrique. Pour la FDA, l'approbation du BiDil® constitue ainsi un « pas vers les promesses de la médecine personnalisée ». Les enjeux d'une telle décision sont extrêmement conséquents car elle sous-entend une reconnaissance officielle d'une logique raciale de la médecine et de la thérapeutique, et promeut l'idée que les différences « raciales » en matière de santé sont essentiellement attribuables à des causes biologiques (Duster, 2007). Il n'y a en effet aucune raison de supposer *a priori* que les différences observées sont dues à des facteurs biologiques et génétiques et non à des facteurs socio-culturels. Si les auteurs avaient pu démontrer l'implication de facteurs génétiques, leur identification dans le génome aurait pu permettre le développement de tests génétiques permettant de conditionner la prescription de ce médicament à la présence des variants recherchés au niveau individuel, plutôt que sur la seule base de l'origine ethnique. N'ayant été testé que chez des patients Africains-Américains, on ne peut pas exclure en effet l'efficacité de ce médicament chez des sujets d'autres origines ethniques (Bloche, 2006). Par ailleurs il ne tient pas compte de l'importante

diversité « inter-raciale » génétique, retrouvée particulièrement chez les individus d'origine africaine (Hoover, 2007; Pena, 2011), ni du fait que la définition de l'identité ethnique ou raciale d'un individu n'est ni clairement définie et standardisée, ni fiable, à tel point que l'auto-identification de l'appartenance ethnique des individus peu varier dans le temps (Krimsky, 2012). Par manque de preuves cliniques réelles que le médicament fonctionnait mieux sur les Africains-Américains, il a été retiré du marché.

L'exemple du BiDil® illustre bien le fait qu'une médecine et une thérapeutique uniquement fondées sur des différences ethniques entre les individus, ne sont ni pertinentes ni valides pour traiter de façon adaptée les différences de réponse aux médicaments observées entre les populations humaines. Certes, l'origine ethnique d'un individu peut donner une indication sur sa probabilité de répondre favorablement à un médicament, mais elle ne peut en aucun cas le prédire de façon certaine (Holm, 2008; Pena, 2011).

4.2 Sources de la variabilité inter-populationnelle dans la réponse aux médicaments

L'origine ethnique n'est donc qu'un indicateur imprécis et inefficace de la réponse aux médicaments, et ne permet pas de guider sans risque la prescription de traitements médicamenteux. En effet, il est important de tenir compte de l'ensemble des facteurs intervenant dans la variabilité de réponse aux médicaments entre les populations humaines, telles que les différences socio-économiques et culturelles. A l'instar de la variabilité interindividuelle de réponse aux thérapies médicamenteuses, cette composante inter-populationnelle peut être déterminée par des facteurs non génétiques et génétiques, avec un rôle particulièrement important des variables agissant au niveau populationnel.

Facteurs non génétiques

Un certain nombre de comportements socio-culturels et de facteurs environnementaux spécifiques de populations comme le mode de vie, les habitudes alimentaires locales, la consommation de tabac et/ou d'alcool, le recours à des techniques de médecines traditionnelles et de suppléments à base de plantes, la présence d'une molécule dans l'environnement

chimique d'une population, peuvent influer sur la pharmacocinétique (biodisponibilité et métabolisme) et/ou la pharmacodynamie (compétition au niveau des récepteurs par exemple) des médicaments (Huang and Temple, 2008; Yasuda et al., 2008). Des différences entre populations peuvent également être liées à des disparités dans le système de soin. Ces dernières sont retrouvées à trois niveaux distincts (Alicea-Alvarez et al., 2013) :

(1) au niveau national, elles concernent l'organisation du système de santé du pays (couverture universelle), les conditions socioéconomiques des individus (influant par exemple sur l'accès à des assurances privées), l'éducation thérapeutique des populations et les différences de disponibilité de certains médicaments et de doses autorisées pour une même molécule entre les pays (Huang and Temple, 2008; Saijo, 2013).

(2) au niveau de la prescription, qui peut être déterminée par des facteurs culturels (pratiques médicales spécifiques de populations, habitudes de prescriptions), mais aussi par la connaissance personnelle du médecin des recommandations existantes et son choix à les suivre.

(3) au niveau du patient, dont l'adhérence au traitement peut être modulée par des facteurs culturels, voire religieux, tels que le langage, les comportements, les systèmes de croyances, l'existence de remèdes populaires,...

Facteurs génétiques

La répartition inégale des variants pharmacogénétiques entre populations humaines peut également conduire à des différences marquées dans la réponse aux médicaments (Yasuda et al., 2008). Celle-ci peut résulter de processus non sélectifs, comme la dérive génétique, les effets fondateurs ou les migrations. Il a été montré par exemple que l'intense processus de dérive génétique qui a affecté certaines populations finlandaises (du fait d'effets fondateurs qui ont eu lieu durant la colonisation de la Finlande et de petites tailles efficaces de populations) a généré une différenciation génétique non négligeable pour les gènes codant pour les cytochromes de la famille CYP2C entre les différentes populations au sein d'un même pays (Sistonen et al., 2009).

Il peut également s'agir de processus sélectifs, avec une action ciblée de la sélection naturelle sur certains gènes de la réponse aux xénobiotiques dans des populations particulières. En effet, du fait de leur rôle de médiateurs entre l'organisme et l'environnement, les gènes de la réponse aux xénobiotiques sont susceptibles de faire l'objet de pressions de sélection naturelle afin de permettre l'adaptation des populations humaines à des environnements chimiques variables dans le temps et dans l'espace. En modifiant le profil de diversité génétique des gènes ciblés par les pressions de sélection dans la ou les population(s) concernée(s), ces événements sélectifs ont pu contribuer à augmenter le niveau de différenciation géographique de certains variants intervenant dans la réponse aux xénobiotiques, dont font partie les médicaments. Aujourd'hui, cette différenciation inter-populationnelle peut se traduire par des différences dans les phénotypes de réponse aux médicaments entre les populations humaines (Li et al., 2011). Les processus sélectifs mis en jeu ainsi que les signatures qu'ils laissent dans le génome seront expliqués en détail dans le chapitre 2 de cette partie 1.

4.3 Les conséquences de cette variabilité

Pour de nombreux variants d'intérêt en pharmacogénétique, une distribution hétérogène entre les populations humaines a été décrite, pouvant se traduire par des différences dans les phénotypes de réponse aux médicaments, aussi bien en terme d'efficacité thérapeutique que de survenue de réactions indésirables. Ces variants peuvent être situés dans les gènes impliqués à tous les niveaux de la réponse aux médicaments (métabolisme, transport, récepteurs cibles, système HLA...). Citons par exemple :

- Métabolisme de phase I : Le CYP2D6 est impliqué dans le métabolisme de plus de 200 médicaments, avec une variabilité dans la réponse qui peut varier d'un facteur 200 (Zanger and Schwab, 2013; Zanger et al., 2004). Les allèles de *CYP2D6* conférant un phénotype de métaboliseur lent présentent une répartition mondiale très inégale : par exemple, l'allèle *CYP2D6*17* est retrouvé presque exclusivement dans les populations africaines, à des fréquences variables au sein de ces populations (9-34 %). L'allèle *CYP2D6*10*

est observé à des fréquences élevées dans les populations asiatiques et faibles dans les populations européennes et africaines, tandis que l'allèle *CYP2D6*4* est principalement retrouvé dans les populations d'origine européenne (Man et al., 2010) (Xie et al., 2001).

- Métabolisme de phase II : les allèles non fonctionnels de *TPMT* affichent une distribution hétérogène entre les populations humaines : l'allèle *TPMT*2* est surtout retrouvé dans les populations africaines, l'allèle *TPMT*3A* dans les populations européennes, *TPMT*3C* dans les populations africaines et coréennes (Man et al., 2010).

- Transport : l'allèle perte de fonction *SLCO1B1*5*, associé au risque de myopathies induites par les statines, est fréquemment retrouvé dans les populations européennes et asiatiques mais très peu dans les populations africaines (Man et al., 2010).

- Pharmacodynamie : L'importante disparité de réponse aux bêta-bloquants entre les patients de d'origine ethnique différente (Materson et al., 1993) pourrait s'expliquer en partie par la distribution hétérogène entre les populations humaines de variants fonctionnels situés dans les gènes codant pour les récepteurs adrénergiques ß1et ß2 (Muszkat, 2007).

- Gènes du système HLA : l'allèle *HLA-B*1502* est considéré comme un prédicteur fiable des réactions d'hypersensibilité graves (pouvant mener à des effets cutanés sévères comme le syndrome de Lyell ou de Stevens-Johnson) induites par la carbamazépine, un épileptique (Chung et al., 2004). Alors que la prévalence de cet allèle est très faible dans les populations européennes et africaines, il est retrouvé plus fréquemment (environ 10 %) dans de nombreuses populations d'Asie (Chine, Singapour, Malaisie, Indonésie, Philippines, Inde, Thaïlande, et Vietnam (Franciotta et al., 2009, Ferrell and McLeod, 2008). Depuis fin 2007, la FDA recommande l'usage d'un test de génotypage pour l'allèle *HLA-B*1502* chez les individus originaires de populations asiatiques avant d'initier un traitement par la carbamazépine (Ferrell and McLeod, 2008). Si le test révèle qu'ils sont porteurs de cet allèle, cette molécule doit être évitée.

4.4 De l'importance de connaître la diversité pharmacogénétique des populations humaines

Insistons sur le fait que, comme évoqué dans l'exemple du BiDil®, le critère de l'origine ethnique pour optimiser les traitements médicamenteux n'est pas suffisamment fiable et précis pour servir de substitut au génotype de l'individu (Yen-Revollo et al., 2008). En effet, différentes études ont montré que les grands groupes habituellement considérés pour distinguer les individus selon leur origine ethnique (par exemple, les africains, européens, asiatiques, ou hispaniques) ne permettent pas de fournir une description efficace de la distribution des polymorphismes des gènes de la réponse aux médicaments, aussi bien à une échelle continentale (Wilson et al., 2001) que nationale (Suarez-Kurtz et al., 2012d). Il est donc nécessaire d'inférer correctement et précisément la structure génétique des populations pour les gènes impliqués dans la réponse aux médicaments afin de définir le niveau approprié de regroupement génétique des populations en sous-groupes génétiquement homogènes pour la réponse clinique aux médicaments. Cela ne peut se faire sans la description de la distribution des variants de la pharmacogénétique dans les populations humaines à une échelle géographique fine.

Les données de populations en pharmacogénétique peuvent néanmoins avoir une utilité à différents niveaux pour guider les autorités compétentes dans la formulation de leurs directives visant à améliorer la prise en charge thérapeutique et lutter contre la iatrogénie médicamenteuse (Ramos et al., 2013). Elles peuvent permettre par exemple de :

- Recommander l'usage des tests pharmacogénétiques à certains individus d'une origine ethnique particulière. Par exemple la FDA préconise de tester l'allèle *HLA-B*1502*, qui augmente le risque de toxicité cutanée par la carbamazépine, uniquement chez les individus d'origine asiatique, ce variant étant principalement retrouvé dans les populations d'Asie (Ferrell and McLeod, 2008).
- Guider le choix des variants génétiques à inclure dans les tests pharmacogénétiques en fonction de la population concernée. En effet, l'amélioration de la puissance des algorithmes pharmacogénétiques pourrait être augmentée dans certaines populations par l'ajout de variants

génétiques principalement observés dans ces populations. Donnons l'exemple des algorithmes de prédiction de la dose thérapeutique de warfarine, un anticoagulant oral de type antivitamine K (AVK). La plus faible performance de ces derniers dans les populations d'origine africaine (Limdi et al., 2008a, 2010) s'explique notamment par la fréquence plus faible en Afrique des principaux variants génétiques inclus dans les modèles de prédiction (Ross et al., 2010), à savoir les variants rs179983, rs1057910 et rs9923231 situés dans les gènes *CYP2C9* et *VKORC1* (*vitamin K epoxide reductase complex, subunit 1*), qui affectent respectivement la pharmacocinétique et la pharmacodynamie des AVK (Takahashi et al., 2006)[5]. Récemment, il a été découvert que l'inclusion du variant rs12777823, situé dans le cluster *CYP2C*, augmentait de 5 % la puissance de l'algorithme élaboré par le IWPC (*International Warfarin Pharmacogenetics Consortium*) chez des individus Africains-Américains (Perera et al., 2013). Or, ce variant n'est pas retrouvé associé avec la réponse à cette molécule chez des individus d'origine européenne, japonaise, ou égyptienne (Perera et al., 2013). Il a également été montré que l'inclusion du variant rs2108622 de *CYP4F2* augmentait la performance des algorithmes pharmacogénétiques de prédiction de la dose de warfarine dans une population chinoise (Liu et al., 2012b), et non dans une autre composée d'individus d'origine ethnique diverse (Lubitz et al., 2010). Cette différence s'explique par la différence de fréquence de l'allèle dérivé de ce variant entre l'Asie et les autres continents (*cf.* Figure 2.5).

- Éclairer les politiques nationales sur l'utilité et l'impact clinique attendu de certains tests pharmacogénétiques en fonction des principaux groupes ethniques d'un pays. Par exemple, la *Health Sciences Authority* (HSA) de Singapour a mesuré les différences de risque de neutropénie induite par l'irinotécan dues à une hétérogénéité de distribution des allèles *UGT1A1*6* et **28* entre les principaux groupes ethniques présents à Singapour (Chinois, Malais et Indiens) (Sung et al., 2011). En observant également que ce risque était augmenté chez ces trois groupes ethniques par rapport aux populations d'origine européenne, une demande de révision de la notice de l'irinotécan par le fabricant a été formulée.

[5] Pour plus de détails, voir la partie 2.

Ces données peuvent également être utiles pour tenir compte de la structure génétique des populations dans les études d'association en pharmacogénétique, où des faux signaux d'association avec le phénotype d'intérêt peuvent être détectés du fait de la présence d'une stratification de population (Delser and Fuselli, 2013).

L'optimisation de la stratégie thérapeutique ne peut donc se faire sans tenir compte de cette composante inter-populationnelle dans la variabilité de la réponse aux médicaments. De nombreux facteurs, aussi bien extrinsèques qu'intrinsèques, interviennent dans cette variabilité (Huang and Temple, 2008). Si l'on peut démontrer que des facteurs génétiques sont en cause, l'approche préconisée en pharmacogénétique consiste à les rechercher dans le génome, en mettant en œuvre des études de type GWAS par exemple. Une fois identifiés, le ou les variant(s) en cause peuvent alors faire l'objet d'un test génétique individuel utilisé en pratique médicale, permettant d'adapter le traitement médicamenteux en fonction de la présence de ce ou ces variant(s). Il peut également être utile d'étudier la distribution dans les populations humaines de variants fonctionnels connus en pharmacogénétique. Les variants présentant un profil de différenciation géographique particulièrement marqué pourront faire l'objet d'une attention particulière, du fait des différences de réponse à certains médicaments qu'ils sont susceptibles d'entraîner, aussi bien en termes d'efficacité que de toxicité. C'est par exemple le cas de nombreux variants situés dans les gènes codant pour les CYP450, dont la distribution hétérogène dans les populations humaines explique une partie de la variabilité inter-populationnelle observée au niveau des phénotypes métaboliques pour un grand nombre de molécules importantes en clinique (McGraw and Waller, 2012). Différentes mesures pourront alors être adoptées au niveau national ou régional pour adapter le choix de la molécule et/ou de la posologie en fonction de la population, ou encore pour recommander ou non l'identification de certains variants génétiques dans le cadre d'un test de génotypage. Cette problématique sera développée dans la partie 3 de cette thèse.

Chapitre 2

Apport de la génétique des populations à la pharmacogénétique

1. Introduction au génome humain

1.1 Rappel sur la variabilité du génome humain

En 1953 a eu lieu la célèbre proposition de la structure de l'acide désoxyribonucléique (ADN) par Watson & Crick (WATSON and CRICK, 1953). Sur la base de cette découverte marquante, la génétique et la biologie moléculaire ont pris leur essor et ont abouti, 50 ans plus tard, au déchiffrage des séquences complètes d'ADN (l'enchaînement exact des nucléotides) notamment celle de l'homme, dont la séquence brute est parue en 2001 (Lander et al., 2001; Venter et al., 2001).

Organisation du génome humain

Le génome humain est le support de l'hérédité génétique des traits biologiques et contient l'ensemble des informations nécessaires au développement et fonctionnement d'un être humain. Il est codé par l'ADN nucléaire (3,1 milliards de paires de bases - pb), divisé en 23 paires de chromosomes (l'homme est une espèce diploïde), fournissant la grande majorité de l'information génétique, et par le petit ADN mitochondrial (16 600 pb).

La molécule d'ADN a une structure spatiale en « double hélice » dans laquelle deux longues chaînes d'acides nucléiques, composées d'un assemblage de nucléotides, sont reliées par des ponts hydrogènes spécifiques. Il existe quatre types de bases composant les nucléotides :

l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T), qui s'apparient de façon complémentaire : ainsi l'adénine établit des liaisons hydrogènes avec la thymine et la cytosine avec la guanine.

Il est aujourd'hui possible d'avoir une représentation globale de la structure du génome humain depuis le succès, 12 années après son lancement en 1989, du projet de séquençage complet de l'ADN du génome humain (Lander et al., 2001). L'ADN nucléaire humain compte deux fois trois milliards de paires de bases. On distingue les séquences codantes, qui peuvent être transcrites en ARN puis traduites en protéines, qui représentent moins de 2 % du génome ; des séquences non-codantes qui ne sont pas transcrites en ARN, ou le sont sans être traduites en protéines. A l'heure actuelle, 20 687 gènes codant pour des protéines sont identifiés (Pennisi, 2012).

Le polymorphisme génétique

Sur la totalité des paires de bases du génome humain, trois à quatre millions diffèrent entre deux individus pris au hasard. Autrement dit, plus de 99,5 % de la séquence d'ADN est identique entre deux individus (Levy et al., 2007). Cependant le développement des techniques de biologie moléculaire a permis de mettre en évidence l'existence d'un polymorphisme génétique correspondant à différentes formes de la séquence d'ADN à certaines localisations spécifiques du génome (locus). Ces formes possibles sont appelées des allèles et la combinaison d'allèles observée à un locus chez un individu constitue son génotype, qui peut être hétérozygote ou homozygote si les deux allèles sont différents ou identiques, respectivement. Cette variabilité génétique retrouvée entre les individus et les populations est causée par les mutations ou les recombinaisons à certains locus, qui modifient à la fois la structure de l'ADN et l'expression des gènes selon les tissus et les organes. Le plus souvent, la cellule répare les dommages de l'ADN mais parfois, les mutations échappent aux mécanismes de réparation. Elles peuvent être transmises à la descendance si elles se produisent dans les cellules germinales et être à l'origine des maladies génétiques simples et complexes et de la variation phénotypique entre les individus et les populations.

On distingue plusieurs grands types de polymorphismes :

- Le polymorphisme chromosomique, provoqué par des réarrangements de segments génomiques allant de quelques centaines de paires de bases à des chromosomes entiers, résultant de différents mécanismes (translocation, inversion, insertions/délétions, duplication).
- Le polymorphisme d'insertion, qui correspond à des éléments d'ADN mobile (transposons) répétitif qui s'insèrent dans le génome par rétrotransposition. Chez les mammifères, ces séquences sont divisées en deux principaux groupes: *long interspersed nuclear elements* (LINE) (6 – 7 kb) et les *short interspersed nuclear elements* (SINE) (< 500 pb), qui représentent 21 et 11 % du génome humain, respectivement.
- Le polymorphisme de répétition, qui correspond à une répétition en nombre variable d'une même séquence de nucléotides : microsatellites (motif de moins de 10 pb), minisatellites (10 à 60 pb), *copy number variation* (CNV) (> 1 000 pb). Ces derniers font partie de la variation structurale du génome.
- Le polymorphisme de substitution, qui correspond au remplacement d'un nucléotide par un autre appelé *single nucleotide polymorphism* (SNP). Cette classe est la plus abondante dans le génome humain : 1 SNP est retrouvé en moyenne tous les 100 à 1000 nucléotides. Ils représentent plus de 90 % de toutes les différences entre individus. Leur répartition sur l'ensemble du génome et la diminution du coût et du temps pour leur identification par génotypage ou séquençage en font d'excellents marqueurs pour l'étude du génome humain. C'est cette dernière catégorie de variation génétique que nous avons considérée dans cette thèse. Notons que selon la fréquence d'occurrence de la substitution nucléotidique, on parlera plutôt de SNV (*single nucleotide variant*) ou de SNP. Les SNVs représentent des variants génétiques nouvellement identifiés, chez un petit nombre d'individus, voire un seul, qui ne sont souvent pas encore validés et répertoriés dans les bases de données.

Les variations génétiques de petite taille, incluant les SNPs, les microsatellites, et les courtes insertions/délétions, sont répertoriées dans la base de données publiques dbSNP, gérée par le NCBI (*National Center for Biotechnology*

Information)(Sherry et al., 2001). Au 23 Juillet 2013, dbSNP comptait 62 676 337 de ces variations dans le génome humain (NCBI, 2013).

Conséquences des mutations sur le phénotype

Les variations génétiques, générées par le processus de mutation, sont présentes dans l'ensemble du génome humain : aussi bien dans les séquences codantes et non codantes des gènes que dans les régions intergéniques. La majorité des mutations sont neutres, c'est-à-dire qu'elles n'apportent ni avantage ni désavantage à celui qui les porte. Les mutations dans une séquence codante n'affectent pas nécessairement la séquence d'acides aminés de la protéine qui est produite du fait de la redondance du code génétique. On parle alors de mutations synonymes. En revanches, ces mutations peuvent avoir un impact sur les phénotypes de différentes manières, en modifiant la transcription, l'épissage, le transport des ARNm, la traduction,… A l'inverse, les mutations non-synonymes (faux sens et non sens) modifient la séquence protéique, et peuvent de ce fait avoir des conséquences plus ou moins importantes sur le phénotype. Quelle que soit leur localisation, les mutations peuvent modifier le phénotype : on parle alors de mutation fonctionnelle. Il faut noter que les relations génotype-phénotype sont extrêmement complexes du fait du grand nombre de facteurs non génétiques pouvant interférer dans la détermination du phénotype. Notons qu'un phénotype peut être déterminé par plusieurs génotypes en interaction (épistasie), qu'un même génotype peut intervenir dans différents phénotypes (pléiotropie) et qu'il existe par ailleurs des interactions avec des facteurs environnementaux et des mécanismes de régulation épigénétique ; leur découverte pour expliquer la base génétique des traits complexes ne fait que commencer.

2. Les différentes forces évolutives et leurs impacts sur la diversité génétique

2.1 La diversité génétique des populations humaines

L'espèce humaine présente un faible niveau de diversité génétique relativement aux autres espèces de grands singes (Kaessmann et al., 2001). Elle est structurée de telle manière que 85 % de la diversité génétique totale est due à des différences individuelles au sein des populations tandis que les 15 % restants séparent entre elles les populations (Lewontin 1972). Cette information, signifiant qu'il n'y a pas beaucoup plus de différences génétiques entre deux individus issus de continents différents qu'entre deux individus vivant dans la même population, a eu un retentissement considérable car elle anéantit le concept de race biologique pour l'espèce humaine.

Il est néanmoins possible à l'aide de marqueurs génétiques, de regrouper les individus en groupes distincts correspondant approximativement à l'appartenance géographique, notamment continentale des individus (Li et al., 2008b; Rosenberg et al., 2002, 2005). En considérant un très grand nombre de marqueurs, des études réalisées en Europe ont réussi à retracer l'origine géographique des individus à une échelle fine (Lao et al., 2008; Novembre et al., 2008). Il est donc pertinent de tenir compte de la population d'origine d'un individu dans les études génétiques, car l'origine géographique est un facteur déterminant une partie de la diversité génétique d'un individu. Nous verrons au cours de cette thèse que cela est vrai pour la variabilité de la réponse aux médicaments.

La génétique des populations a pour objectif de décrire, évaluer et interpréter le rôle respectif des différentes forces évolutives qui sont à l'origine de la diversité génétique des populations d'une même espèce. Celles-ci peuvent agir à différents niveaux : génomique (mutations et recombinaisons), démographique (taille des populations, migrations), et sélectif (environnements climatiques, pathogéniques, nutritionnels). Ces forces ont

agi avec une intensité variable dans le temps et dans l'espace et ont façonné la diversité génétique au niveau populationnel. La diversité génétique varie selon les populations et la portion du génome considérée.

L'exploration du rôle joué par ces différentes forces évolutives dans les profils de diversité génétique observés aujourd'hui au sein et entre les populations humaines permet de mieux comprendre les mécanismes moléculaires en jeu dans l'adaptation de l'homme à son environnement, à l'origine de la grande diversité phénotypique des populations humaines. A une échelle globale en effet, les populations humaines affichent une grande hétérogénéité phénotypique (couleur de la peau, forme du visage et du corps, …), ainsi qu'une susceptibilité génétique aux maladies différentielle (par exemple la susceptibilité aux agents infectieux) (Lewontin 1995). En étudiant la diversité génétique des populations humaines et le poids des forces évolutives sous-jacentes, les approches de génétique évolutive et des populations sont d'une grande utilité en génétique médicale. Elles permettent de mieux comprendre la transmission et l'épidémiologie des maladies génétiques, et d'expliquer le maintien de certains allèles morbides (notamment pourquoi certaines maladies très défavorables se maintiennent à une forte fréquence dans certaines populations) et, on le verra au cours de cette thèse, elles peuvent faciliter l'identification de gènes ou de variants impliqués dans le déterminisme de traits complexes comme la réponse aux médicaments.
Par ailleurs, il faut noter qu'en plus des ces forces évolutives que nous allons à présent décrire, l'organisation sociale et les habitudes culturelles complexes des êtres humains peuvent également influencer la structure génétique des populations (Chaix et al., 2007).

2.2 Les forces génomiques

La mutation et la recombinaison sont les sources classiques de la variation biologique. Les mutations apparaissent de manière aléatoire sur le génome. Elles sont soit endogènes (produit des erreurs spontanées de réplication de l'ADN), soit causées par l'exposition à des agents mutagènes physiques (rayonnements ionisants, UV) ou chimiques. Une mutation dans une cellule

germinale peut se transmettre à la descendance. Le taux de mutation varie selon le type de mutation et le locus considéré. En moyenne, il est estimé pour le génome humain à environ 2,5 x 10^{-8} par nucléotide, ce qui correspond à environ 160 mutations par génome diploïde par génération (Nachman and Crowell, 2000). Chaque nouvelle mutation est liée génétiquement aux autres mutations voisines déjà présentes. Les allèles observés ensemble à des loci adjacents sur un même chromosome sont en déséquilibre de liaison (DL) et leur combinaison forme un haplotype. La recombinaison engendre des nouvelles combinaisons alléliques, et donc de nouveaux haplotypes, par brassage intrachromosomique ou interchromosomique du matériel génétique. Mutations et recombinaison créent de la diversité génétique en générant de nouveaux allèles et haplotypes. D'autres forces évolutives, comme la dérive génétique, la migration et la sélection naturelle, vont ensuite agir sur ces nouveaux variants et modifier leur fréquence et leur distribution dans les populations humaines.

2.3 La dérive génétique

La dérive génétique est la fluctuation aléatoire des fréquences alléliques dans une population de taille finie au cours des générations (Wright, 1931). En l'absence de sélection, la transmission des gamètes se fait au hasard, conduisant à une variation des fréquences alléliques d'une génération à l'autre. Les effets de la dérive génétique dépendent de la taille de la population. Dans les populations de grand effectif, les fréquences alléliques restent stables durant plusieurs générations successives. A l'inverse, les écarts de fréquences alléliques d'une génération à l'autre vont être d'autant plus grands que la taille de la population est faible. Ce phénomène concerne surtout les allèles neutres ne conférant ni avantage, ni désavantage sélectif. Le processus de dérive génétique conduit inexorablement à la perte de certains allèles et à la fixation de certains autres, contribuant à la réduction de la diversité génétique au sein d'une population par la perte progressive du polymorphisme génétique.

L'effet fondateur, cas particulier de dérive génétique, correspond à la réduction de la diversité génétique dans une population ayant pour origine un petit nombre d'individus (population fondatrice) issus d'une population mère (Slatkin and Excoffier, 2012). Ces individus ne vont « emporter » qu'un échantillon d'allèles du pool d'allèles de la population mère, et ce de manière que l'on suppose aléatoire. La nouvelle population peut donc présenter des fréquences génotypiques fort différentes de la population initiale. Cet écart peut changer radicalement le profil (allélique, génotypique et phénotypique) de la population fondatrice, par rapport à la population initiale. Cet effet est responsable de la fréquence élevée de certaines maladies génétiques rares dans certaines populations, en raison de l'augmentation en fréquence d'allèles délétères.

2.4 La migration

La migration est l'échange d'individus (ou de gamètes) entre des sous-populations (qui peuvent présenter des différences génétiques), permettant les flux de gènes entre elles. Il peut s'agir d'une fusion de populations ou du mélange de façon unidirectionnelle d'une population de migrants, qui se déplacent dans un nouvel environnement géographique, à une population déjà implantée sur le lieu. Elle conduit à une homogénéisation des fréquences alléliques et à une diminution du niveau de différenciation génétique des deux populations.

La dérive génétique et les migrations ont eu un rôle important dans l'histoire génétique de l'humanité et ont déterminé une grande part des niveaux observés de diversité génétique au sein des populations et des niveaux de différenciation génétique entre les populations. L'étude combinée de données génétiques sur l'ADN mitochondrial (Cann et al., 1987; Ingman et al., 2000), le chromosome Y (Thomson et al., 2000), le chromosome X (Harris and Hey, 1999) et les autosomes (Jorde et al., 1997), et de données archéologiques, paléontologiques et linguistiques (Cann, 2001) ont pu mettre en évidence l'origine africaine de notre espèce. La plupart des études génétiques conduites sur un grand nombre de locus indépendants du

génome ont démontré une plus grande diversité génétique globale des populations africaines par rapport au reste du monde, et que la majeure partie de la variabilité génétique observée dans les populations non africaines représentait un sous-ensemble de celle observée dans les populations africaines. L'étude de la variabilité génétique, haplotypique et des profils de déséquilibre de liaison à l'échelle mondiale ont confirmé cette conclusion (Conrad et al., 2006; Gabriel et al., 2002; Jakobsson et al., 2008; Stephens et al., 2001). Ces observations sont à l'origine du modèle *Out of Africa* (Lewin, 1987), selon lequel un groupe fondateur d'être humains a quitté l'Afrique de l'Est il y a environ 60 000 ans pour coloniser, par vagues de migrations successives, le Proche-Orient, l'Asie d'où a démarré la colonisation de l'Océanie, l'Europe et enfin l'Amérique (Cavalli-Sforza and Feldman, 2003; Henn et al., 2012). Au fil de ces vagues de colonisation, les populations migrantes n'emportaient avec elles qu'une partie de la variabilité génétique de leur population d'origine (effet fondateur) et leur taille diminuant au fur et à mesure, les effets de la dérive ont augmenté (Keinan et al., 2007; Marth et al., 2003). L'isolement progressif des populations par la distance a contribué à cette différenciation génétique et explique le haut degré de corrélation entre les distances génétiques et géographiques séparant les populations humaines (Cavalli-Sforza and Feldman, 2003; Prugnolle et al., 2005).

Les forces génomiques, la dérive génétique et les migrations ont en commun d'affecter le génome dans sa globalité. En cela, elles diffèrent de la sélection naturelle qui a une action plus ciblée sur certaines régions du génome.

2.5 La sélection naturelle

Cette force évolutive est, telle que proposée par Charles Darwin (Darwin, 1859), le mécanisme majeur de transformation et de diversification évolutive des espèces. Elle agit par le tri, au sein d'une population, des individus les mieux adaptés aux contraintes de leur milieu qui, de ce fait, ont une meilleure survie et laissent un plus grand nombre de descendants (*fitness* augmenté). En permettant ainsi la propagation des traits biologiques avantageux dans une population, la sélection naturelle a un rôle

déterminant dans l'adaptation des populations à leur environnement. D'un point de vue génétique, la survie et la fécondité différentielles des individus selon leur génotype (valeur sélective ou coefficient de sélection) modifient au fil des générations la fréquence des variants génétiques sélectionnés et fait, à terme, évoluer la structure génétique d'une population. La sélection naturelle peut, en effet, en quelques générations seulement, faire varier fortement les fréquences alléliques. Elle a pour conséquence la fixation des allèles favorables, et la disparition des allèles qui ne le sont pas, ou encore le maintien d'allèles à une fréquence d'équilibre stable sur une longue période de temps.

Chez l'homme, les signatures génomiques de l'action de la sélection naturelle sont nombreuses. En effet, comme nous l'avons expliqué dans le premier chapitre de cette partie, les populations humaines ont dû faire face au cours de leur histoire évolutive à des variations de leur environnement (climatique, nutritionnel, pathogénique, etc.), à l'origine de pressions de sélection sur le génome (Vasseur and Quintana-Murci, 2013).
L'intérêt de détecter les signatures génomiques de la sélection naturelle est double. Il relève tout d'abord de notre curiosité naturelle à connaître et comprendre l'histoire passée de l'être humain, ainsi que les mécanismes sous-jacents à son évolution. La deuxième motivation concerne l'identification des gènes et des variants ayant une importance fonctionnelle significative pour la survie et la bonne santé de l'être humain, que la sélection a le potentiel de révéler (Nielsen et al., 2007; Sabeti et al., 2006).

Exemples de sélection naturelle chez l'homme

Ainsi, la détection de l'action de la sélection naturelle sur le génome humain a permis d'identifier des gènes fonctionnellement importants ayant contribué au phénomène de spéciation conduisant à l'être humain moderne, et notamment ceux impliqués dans son développement cognitif. C'est le cas du gène *FOXP2 (forkhead box P2)*, qui a été la cible de la sélection positive dans l'espèce humaine, favorisant l'acquisition du langage oral (Enard et al., 2002; Zhang et al., 2002). Des mutations dans ce gène entraînent des troubles sévères du développement de la parole et du langage oral incluant des

difficultés d'articulation, des troubles du langage et des carences grammaticales (Fisher et al., 1998; Lai et al., 2001; Vargha-Khadem et al., 1995).

De nombreux gènes impliqués dans l'adaptation des populations humaines à différents milieux ont également été découverts. Parmi les exemples les plus connus, nous en citerons trois faisant intervenir une pression de sélection relative à l'alimentation, les pathogènes et le climat.

Le premier exemple est celui de la persistance de la production de la lactase (enzyme qui permet de digérer le lactose) à l'âge adulte, liée à certains allèles du gène *LCT* ayant pour effet de maintenir un niveau élevé d'expression du gène et de ce fait, de la synthèse de l'enzyme, au-delà de la période de sevrage (Schlebusch et al., 2013). Une forte pression de sélection positive favorisant ces allèles dans certaines populations humaines a été mise en évidence au locus de ce gène et semble être directement liée au mode de subsistance des populations. Plusieurs mutations différentes auraient été sélectionnées indépendamment en Europe (entre -5000 et -10000 avant JC) et en Afrique (entre -3000 et -7000 avant JC), aboutissant au même résultat de conservation de la lactase à l'âge adulte (adaptation convergente) (Bersaglieri et al., 2004; Ranciaro et al., 2014; Tishkoff et al., 2007). Une corrélation entre une tradition ancienne d'élevage et une fréquence élevée du phénotype de persistance de la lactase a été observée et la période au cours de laquelle les mutations seraient devenues avantageuses semble correspondre au moment de l'essor de « l'industrie » laitière. Les individus maintenant la capacité à digérer le lait à l'âge adulte auraient été avantagés du fait du considérable apport nutritionnel, énergétique et hydrique que le lait pouvait leur apporter. La sélection du trait de persistance de la lactase figure parmi les plus beaux exemples de coévolution gène-culture.

Un deuxième exemple classique, lié cette fois à l'adaptation à l'environnement pathogène, est celui du maintien par la sélection positive à des fréquences élevées de certains variants conférant une déficience en G6PD (glucose-6-phosphate déshydrogénase), enzyme cellulaire ayant un rôle métabolique clé pour le globule rouge. En dépit des troubles

hématologiques (anémies hémolytiques) qu'une faible activité de cette enzyme peut entraîner, ces variants confèrent une résistance significative au paludisme, et ce, dans différentes régions du monde (Louicharoen et al., 2009; Ruwende et al., 1995; Tishkoff et al., 2001).

Enfin, on peut citer l'adaptation locale à la vie en haute altitude (> 2 500 mètres au-dessus du niveau de la mer) comme exemple de réponse à une pression de sélection climatique. Des signatures de sélection positive ont en effet été détectées dans les populations andéennes et tibétaines pour des gènes permettant une meilleure utilisation de l'oxygène (Bigham et al., 2010). C'est le cas notamment du gène *EPAS1* (*endothelial PAS domain protein 1*), facteur de transcription impliqué dans la réponse à l'hypoxie (baisse de concentration en oxygène). Les écarts de fréquences particulièrement importants observés pour un variant de ce gène entre les populations tibétaines et chinoises en font le changement de fréquence allélique le plus rapide rapporté à l'heure actuelle (Yi et al., 2010).

Dans un second temps, étant donné que la sélection naturelle cible des variants fonctionnels déterminant des phénotypes (Nielsen et al., 2007), l'exploration des signatures génomiques de la sélection peut permettre de mettre en évidence des gènes, voire des variants génétiques causaux, impliqués dans des maladies ou d'autres traits phénotypiques d'intérêt chez l'homme. C'est ainsi qu'un enrichissement en signatures de sélection positive a été mis en évidence pour les gènes impliqués dans la réponse immunitaire et dans la susceptibilité aux maladies infectieuses (Barreiro and Quintana-Murci, 2010; Casto and Feldman, 2011; Raj et al., 2013). Ces observations concordent avec l'hypothèse dite « hygiéniste » (*hygiene hypothesis*), qui propose qu'une réponse immunitaire très développée a été probablement favorisée dans le passé pour lutter contre des agressions pathogéniques multiples et variées. Or, depuis l'amélioration dans la vie moderne des conditions d'hygiène et l'utilisation des antibiotiques, le risque d'infection est moindre et ces génotypes de résistance aux pathogènes seraient devenus délétères, favorisant l'essor des maladies auto-immunes, inflammatoires et allergiques (Strachan, 2000).

Ces différents exemples illustrent à quel point l'identification des gènes ciblés par la sélection peut permettre d'améliorer notre compréhension de la base génétique de traits d'intérêt chez l'homme. Cela est particulièrement vrai pour les phénotypes d'importance en génétique médicale, les variants génétiques conférant une plus grande susceptibilité ou résistance à des maladies étant de particulièrement bons candidats pour la sélection naturelle.

La sélection naturelle peut prendre différentes formes et agir avec une intensité variable, ce que nous nous proposons de décrire à présent.

3. Les différentes formes de sélection naturelle et la détection de leurs signatures moléculaires

3.1 La théorie neutraliste de l'évolution

La théorie neutraliste de l'évolution moléculaire décrite par Motoo Kimura en 1968 pour expliquer la diversité des populations d'un point de vue moléculaire, postule que la plupart des mutations génétiques sont neutres au regard de la sélection naturelle et que la fluctuation de leurs fréquences alléliques au sein des différentes populations est majoritairement déterminée par l'action aléatoire de la dérive génétique (Kimura, 1968). Sans renier le rôle de la sélection naturelle dans les processus adaptatifs populationnels, la dérive génétique devient alors la principale pression évolutive déterminant la différenciation génétique des populations. Cette théorie constitue actuellement un pan entier de la théorie synthétique de l'évolution. Elle est à la base des tests de détection de la sélection naturelle.

3.2 Les différentes formes de sélection naturelle

On distingue trois grands types de sélection au niveau moléculaire.

a. La sélection purificatrice

Aussi appelée sélection négative ou sélection d'arrière plan, elle consiste en la réduction de la fréquence des allèles délétères (qui confèrent un désavantage sélectif aux individus qui les portent) jusqu'à leur élimination complète dans la population. Il s'agit sans doute de la sélection la plus fréquente dans le génome, car elle permet le maintien à long terme des fonctions biologiques essentielles (Bamshad and Wooding, 2003; Nielsen et al., 2007). Les gènes domestiques, ou de ménage, indispensables au fonctionnement de base de la cellule sont vraisemblablement les plus soumis à cette contrainte sélective. Elle a la particularité d'agir sur tout le génome (Charlesworth et al., 1993).

b. La sélection positive

Au contraire, la sélection positive entraîne une augmentation rapide de la fréquence des allèles avantageux (qui confèrent un avantage sélectif pour la survie et la reproduction de l'individu) jusqu'à leur fixation dans la population. Elle agit en générant ce que l'on appelle un balayage sélectif (*selective sweep*), mécanisme décrit par John Maynard Smith & John Haigh en 1974 (Smith and Haigh, 1974), qui correspond à l'enchaînement des allèles situés à proximité du locus sélectionné (effet d'auto-stop génétique), qui se traduit par une forte réduction de la variation génétique avoisinante, sur une étendue qui dépend de la force de la sélection et des taux de recombinaison locaux (Figure 1.6). Dans la région génomique entourant le locus sélectionné, on peut alors observer des déséquilibres de liaison forts et un excès de variants avec des fréquences élevées.

Ces balayages sélectifs peuvent être complets, c'est-à-dire que l'allèle sélectionné est fixé dans la population, ou partiels lorsque la fréquence est élevée sans toutefois atteindre la fixation. Ce dernier cas correspond à une pression de sélection en cours, ou qui s'est affaiblie voire arrêtée avant que l'allèle n'ait eu le temps d'arriver à la fixation (Figure 1.6).

Figure 1.6 | Représentation schématique des différents types de balayages sélectifs. La distribution des mutations (étoiles) est représentée sur les chromosomes (barres grises) avant sélection, et après un balayage sélectif partiel ou complet. Inspiré de Nielsen *et al.* (2007).

Ces scénarios de sélection positive considèrent que la sélection opère sur une mutation nouvelle, et donc par principe rare. D'autres formes de sélection positive agissant sur des allèles déjà présents à une certaine fréquence dans la population ont également été décrites (Hermisson and Pennings, 2005). Du fait de la dérive génétique, ces allèles peuvent être retrouvés à des fréquences variables dans la population. S'ils deviennent avantageux lors d'un changement de conditions environnementales par exemple, la sélection positive peut alors les sélectionner à travers deux configurations possibles : (1) le *soft sweep*, qui décrit le processus par lequel,

à un même locus, plusieurs allèles préexistants deviennent avantageux et augmentent en fréquence de façon concomitante ; (2) l'adaptation polygénique qui cible plusieurs variants à des loci différents en même temps (Hermisson and Pennings, 2005; Messer and Petrov, 2013; Pritchard et al., 2010) (Figure 1.7).

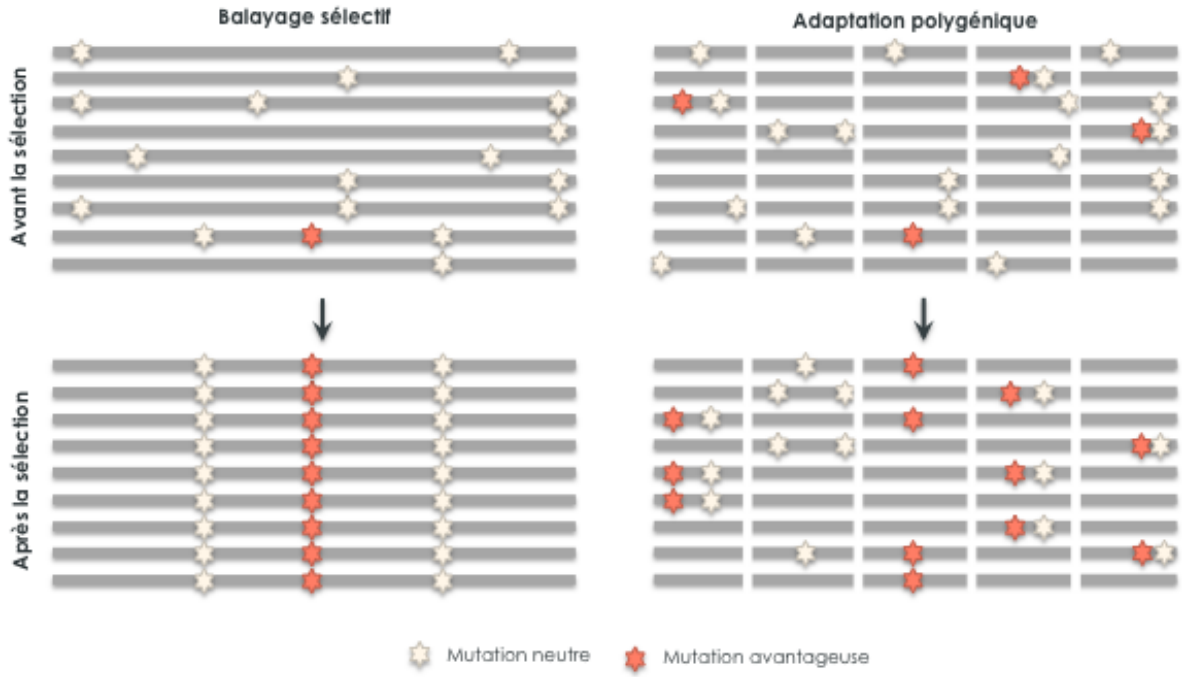


Figure 1.7 | Représentation schématique d'un balayage sélectif classique et de l'adaptation polygénique. La distribution des mutations (étoiles) est représentée sur les chromosomes (barres grises) avant sélection et après sélection positive. Inspiré de (Pritchard et al., 2010).

c. La sélection balancée

Ce régime de sélection se réfère à un certain nombre de processus par lesquels plusieurs allèles à un même locus sont activement maintenus dans une population à des fréquences intermédiaires. La sélection balancée a donc pour effet d'augmenter la variabilité génétique dans la population au locus sélectionné, et de réduire la différenciation génétique des populations (Bamshad and Wooding, 2003). Chez l'homme, les exemples les plus courants concernent des gènes impliqués dans la réponse aux pathogènes. Plusieurs cas peuvent se présenter :

- *L'avantage de l'hétérozygote* qui favorise les individus portant deux allèles différents par rapport aux individus homozygotes. L'exemple le plus connu est le maintien à des fréquences relativement élevées de l'allèle morbide de l'hémoglobine S (allèle HbS) dans les régions du monde où sévit le paludisme (il s'agit d'un cas d'évolution convergente) (Feng et al., 2004). A l'état homozygote, cet allèle est responsable de la drépanocytose (maladie touchant les globules rouges qui se manifeste par une anémie hémolytique, des infections et des crises vaso-occlusives). A l'état hétérozygote, cet allèle permet d'augmenter d'un facteur 10 la résistance au paludisme sévère. Le paludisme représente la pression sélective la plus forte connue sur l'évolution récente du génome humain (Kwiatkowski, 2005).
- *La sélection fréquence-dépendante* correspond à un processus évolutif où l'aptitude d'un phénotype à persister dans le temps dépend de sa fréquence par rapport à d'autres phénotypes dans une population donnée. C'est le cas par exemple des allèles qui sont avantagés quand ils deviennent rares (avantage du rare). Les types rares seront sélectionnés jusqu'à ce qu'ils deviennent les plus abondants, d'où une possible sélection qui varie dans le temps. Les exemples les plus connus concernent ce qui a trait au système immunitaire et à la sélection sexuelle (Schierup et al., 2001). C'est ce qui est observé par exemple pour les gènes du système HLA (*Human Leukocyte Antigen*) de classe I et II codant pour le complexe majeur d'histocompatibilité (MHC). Il a été mis en évidence qu'un couple dont le système HLA est trop similaire est stérile, et qu'il y a manifestement des attirances dépendantes de la différence entre le HLA des deux partenaires (Chaix et al., 2008; Laurent and Chaix, 2012; Wedekind and Penn, 2000).

Les différentes formes que peut prendre la sélection naturelle sont résumées dans la Figure 1.8.

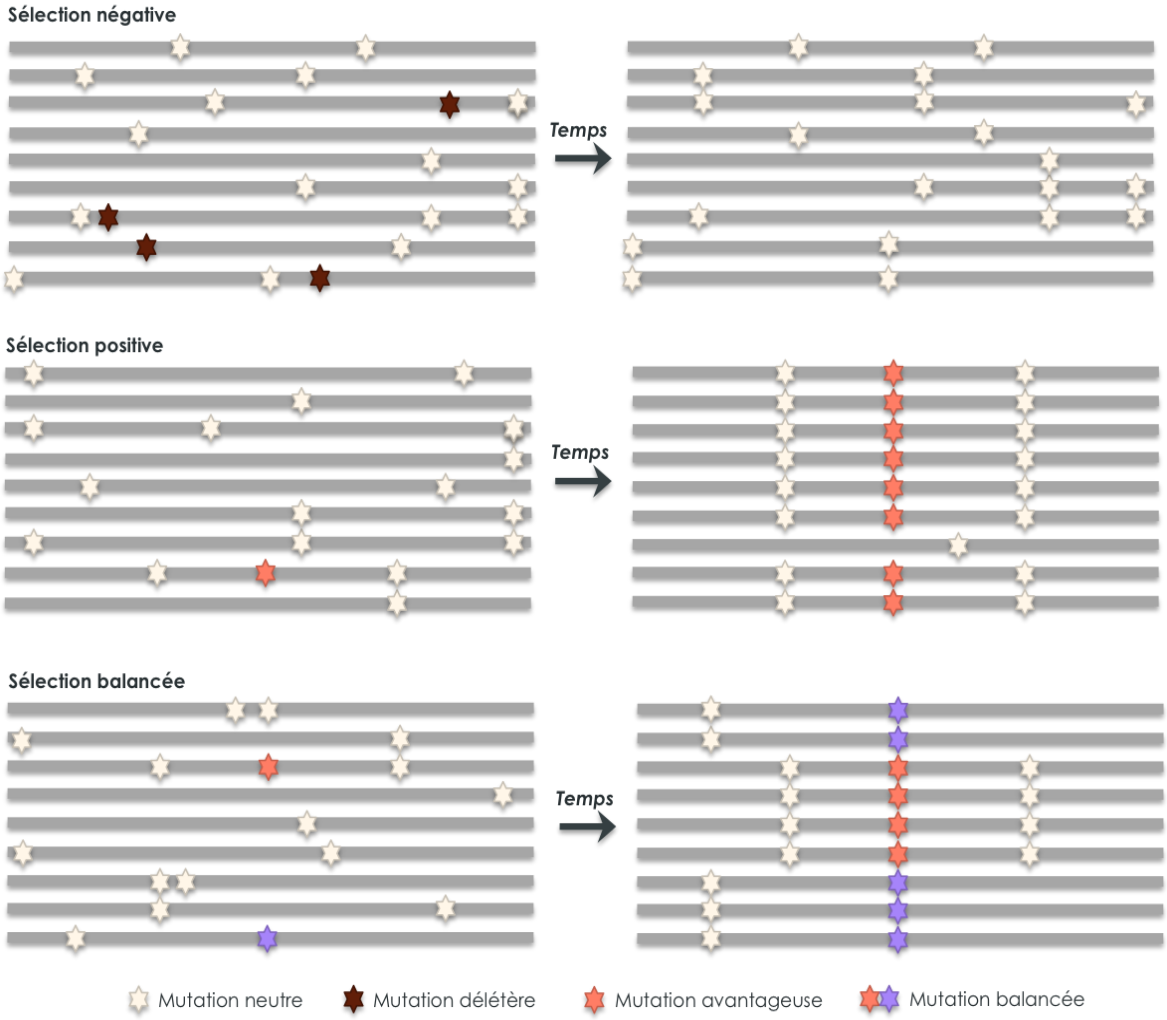


Figure 1.8 | Représentation schématique des différentes formes de sélection naturelle. La distribution des mutations (étoiles) est représentée sur les chromosomes (barres grises) avant sélection et après sélection négative, positive et balancée. Inspiré de (Quintana-Murci and Clark, 2013).

3.3 Tests de détection de la sélection naturelle

Un grand nombre de tests de détection de la sélection naturelle ont été développés, en particulier pour la sélection positive. Nous les décrirons rapidement ici, en détaillant particulièrement ceux que nous avons employés dans nos travaux. Les tests de sélection sont basés sur les différents aspects de la variation génétique modifiés par la sélection, qui perdurent selon des échelles de temps différentes (Sabeti et al., 2006) à savoir :

- La proportion de mutations ayant des effets sur les protéines,

- Le spectre de fréquence allélique (*i.e*. l'histogramme des fréquences alléliques aux locus polymorphes),
- La différenciation génétique des populations,
- La structure haplotypique.

Pour l'ensemble de ces tests, l'absence de sélection constitue l'hypothèse nulle, c'est-à-dire que les variants génétiques ont un effet neutre. Sous cette hypothèse de neutralité, la diversité génétique n'est affectée que par la dérive, les migrations, les mutations et la recombinaison.

Tests de neutralité inter-espèces

La sélection positive augmente le taux de fixation de mutations fonctionnelles favorables. Ces changements sont détectables en comparant la séquence d'ADN entre espèces. Les tests de neutralité inter-spécifiques se basent sur le niveau de divergence (variation inter-espèces) et de polymorphisme (variation intra-espèces) entre les différentes catégories de mutations. Par exemple, un excès du rapport mutations non-synonymes/mutations synonymes chez l'homme par rapport au chimpanzé est un indicateur de la sélection adaptative chez l'homme. Ils permettent donc de détecter de la sélection ancienne.

Tests intra-spécifiques

Les tests intra-spécifiques se basent sur le polymorphisme observé au sein des populations. Ils sont donc capables de détecter des pressions de sélection plus récentes que les tests inter-spécifiques, en rapport avec non pas la spéciation adaptative de l'homme mais son adaptation à différents environnements.

a. Tests basés sur le spectre de fréquence allélique

La sélection modifiant la distribution des fréquences alléliques à l'intérieur d'une population, différents tests statistiques ont été proposés qui évaluent si le spectre de fréquence allélique observé dans la population est conforme à celui attendu sous la neutralité. Citons le *F* et le *D* de Fu & Li (Fu, 1997; Fu and Li, 1993), le *H* de Fay & Wu (Fay and Wu, 2000), le *D* de Tajima (Tajima, 1989a). Nous avons utilisé le *D* de Tajima dans cette thèse, qui mesure la différence entre deux estimateurs du taux de mutation dans la population : le nombre

total de sites polymorphes observés (θw) et la moyenne du nombre de différences observées entre paires de séquences ($\theta \pi$). Sous l'hypothèse de neutralité, ces deux estimateurs doivent être égaux ou très proches et on s'attend donc à ce que la statistique *D* ne soit pas significativement différente de 0. Une valeur négative de *D* indique un excès d'allèles rares et une valeur positive un excès d'allèles à fréquences intermédiaires, caractéristiques de la sélection positive/négative et balancée, respectivement (Figure 1.9). De façon similaire, les statistiques *F* et *D* de Fu & Li mesurent un excès ou déficit de nouveaux allèles, en utilisant chacun un estimateur du taux de mutations basé sur le nombre de singletons. L'interprétation de leur valeur est la même que celle du *D* de Tajima : négative, elle indique de la sélection directionnelle, et positive, de la sélection balancée. La statistique *H* de Fay & Wu, basée sur un estimateur du taux de mutation sensible aux mutations dérivées ayant une fréquence élevée, permet de distinguer la sélection positive de la sélection négative, qui présentent toutes les deux un excès de mutations rares. Elle mesure en effet un excès de mutations dérivées à forte fréquence, ce qui est une caractéristique spécifique aux évènements récents de sélection positive (Fay and Wu, 2000). Ce test requiert d'utiliser une autre espèce proche pour obtenir l'information du statut allélique : ancestral ou dérivé.

Les signaux significatifs détectés avec les tests mesurant un excès d'allèles rares peuvent persister durant plusieurs centaines de milliers d'années, assez longtemps pour remonter aux origines de l'homme moderne. Le *H* de Fay & Wu, qui s'intéresse aux allèles dérivés de fréquence élevée, et détecte des évènements sélectifs plus récents (< 80 000 ans).

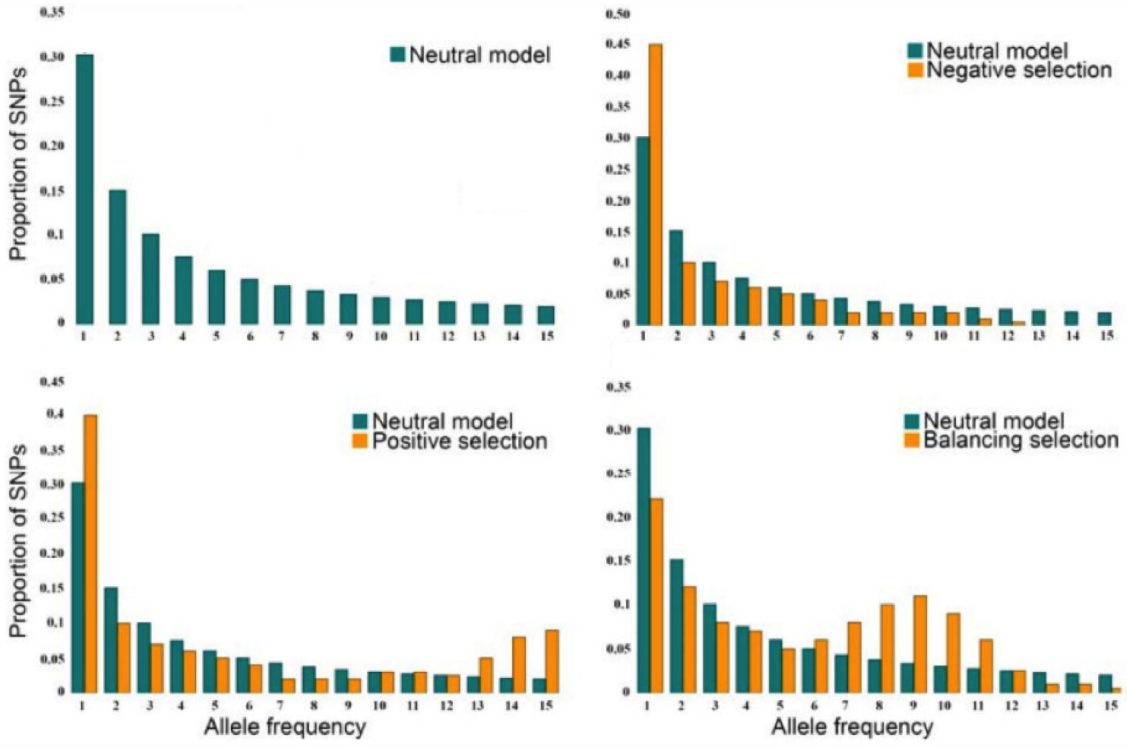


Figure 1.9 | Différentes formes prises par le spectre de fréquence allélique d'une population sous un modèle neutre ou sous sélection négative, positive ou balancée.

b. Tests basés sur la différenciation génétique inter-populationnelle

1. Le F_{ST}

La sélection naturelle peut conduire à des différences de répartition des allèles entre différentes populations et un moyen de la détecter consiste donc à quantifier les différences de fréquences alléliques entre populations. Pour ce faire, une mesure classiquement utilisée est l'indice de fixation F_{ST}. Cet indice, proposé par Wright en 1951, mesure la corrélation entre deux gamètes échantillonnés dans une même sous-population par rapport à l'ensemble de la population (WRIGHT, 1951).

Parmi les différents estimateurs du F_{ST}, nous avons utilisé dans cette thèse celui de Weir et Cockerham (Weir and Cockerham, 1984). Son calcul se base sur une décomposition de la variance totale des fréquences alléliques observées en différentes composantes de la variance selon la formule suivante :

$$FST = \frac{a}{a+b+c}$$

Avec : a = proportion de la variance génétique totale attribuée aux différences entre populations (variance inter-populationnelle),

b = proportion de la variance génétique totale attribuée aux différences entre individus au sein des populations (variance intra-populationnelle),

c = proportion de la variance génétique totale attribuée aux différences entre les gamètes chez les individus (variance intra-individuelle).

L'estimateur de cet indice permet de mesurer le degré de différenciation des populations, puisqu'il estime la part de variance interindividuelle expliquée par la variance entre populations.

Dans nos travaux, nous avons calculé le F_{ST} en regroupant les individus selon différents niveaux de subdivision : par continents (F_{ST} inter-continental) et par populations (F_{ST} inter-populationnel).
La valeur de cette statistique varie de 0, pour des populations génétiquement identiques à 1, pour des populations totalement différenciées génétiquement. Les estimations du F_{ST} réalisées à partir d'un grand nombre de marqueurs génétiques de types SNPs répartis sur l'ensemble du génome ont mis en évidence une différenciation génétique modérée des populations humaines, avec une valeur moyenne de F_{ST} de l'ordre de 0,10-0,15 (Akey et al., 2002; Barbujani and Colonna, 2010; Barreiro et al., 2008; Weir et al., 2005). Ces estimations ont longtemps été basées sur un petit nombre de populations représentant les grandes populations continentales, comme les Yoruba, Chinois, Japonais et Européens du projet HapMap (International HapMap Consortium, 2005). Plus récemment, les estimations de la différenciation génétique moyenne des populations humaines, basées sur un plus grand nombre de populations dans le monde ont révélé des valeurs plus faibles, de l'ordre de 0,05-0,010 (1000 Genomes Project Consortium et al., 2010; Auton et al., 2009; Barbujani and Colonna, 2010).
La valeur du F_{ST} dépend fortement de la fréquence de l'allèle mineur du locus considéré (Beaumont MA, Nichols RA (1996). Par conséquent il est

nécessaire de comparer les valeurs de F_{ST} de variants dont les allèles mineurs ont une fréquence similaire si l'on veut identifier des profils de différenciation génétique atypique pour des variants d'intérêt. C'est ce que nous avons fait dans les travaux présentés dans cette thèse, en calculant les *P*-values des SNPs au sein de classes de MAF (*Minor Allele Frequency*).

Sous la neutralité, la valeur de F_{ST} est influencée par la dérive génétique de façon similaire sur l'ensemble du génome. La sélection naturelle en revanche, va augmenter ou diminuer la différenciation génétique des populations à certains loci spécifiquement. La sélection directionnelle agissant le plus souvent à une échelle géographique locale, c'est-à-dire ne concernant qu'un nombre restreint de populations, a pour effet d'augmenter la différenciation génétique des populations ; on obtient alors des valeurs élevées du F_{ST} au locus sélectionné. A l'inverse, on constate une diminution des valeurs du F_{ST} lorsque la différenciation génétique des populations est réduite, ce qui advient dans le cas de la sélection balancée ou de la sélection directionnelle intervenant dans plusieurs populations en même temps.

Cette statistique présente l'avantage de pouvoir être calculée pour chaque variant génétique et non au sein d'une fenêtre génomique comprenant plusieurs variants, ce qui permet potentiellement d'identifier le variant précis ciblé par la sélection, et du moins d'affiner la localisation spatiale d'un événement sélectif. Étant donné que le F_{ST} mesure les différences de fréquence allélique observées entre populations au locus sélectionné, il permet de détecter des évènements sélectifs relativement récents qui sont apparus une fois les populations humaines isolées d'un point de vue reproductif, autrement dit, après les grandes vagues de migrations des populations (Sabeti et al., 2006). L'échelle de temps varie bien-sûr en fonction des populations considérées mais elle est de l'ordre de quelques dizaines de milliers d'années.

2. Le test XP-CLR

Une approche intéressante basée sur la mesure de la différenciation génétique inter-populationnelle au sein d'une région génomique étendue a été développée récemment avec le test XP-CLR (*Cross population Composite Likelihood Ratio*) (Chen et al., 2010a).

La statistique XP-CLR mesure de façon simultanée la différenciation génétique entre deux populations (une population testée et une population de référence) pour plusieurs allèles d'une région génomique. Si de la sélection positive est intervenue dans la population testée au niveau de la région génomique analysée, une distorsion du spectre de fréquence allélique par rapport à ce qui est attendu sous la neutralité est observée. Cette signature s'atténue au fur et à mesure que l'on s'éloigne du variant sélectionné. De plus, la taille de la région génomique concernée par une importante différenciation allélique inter-populationnelle est fonction de l'intensité de la sélection. Ce sont ces marques visibles sur le génome que la statistique XP-CLR détecte, par une modélisation de la différenciation allélique multilocus de deux populations, qui est fonction de l'intensité de la sélection et de la distance génétique séparant le variant directement ciblé par la sélection et les allèles neutres situés dans la région génomique sous sélection. En conséquence, cette statistique de test se calcule pour une région génomique donnée, et non, comme le F_{ST}, pour un seul variant génétique.

Ce test permet de détecter de façon puissante un balayage sélectif complet ou presque complet. A l'instar du F_{ST}, qui se base sur les différences de fréquence allélique entre populations, il ne peut détecter que des évènements sélectifs relativement récents dans l'histoire évolutive de l'homme.

c. Tests basés sur le niveau de déséquilibre de liaison

Comme nous l'avons expliqué, la sélection positive va amener rapidement un allèle avantageux à une fréquence élevée dans une population en entraînant avec lui la variation génétique adjacente. Cette augmentation rapide de la fréquence allélique du variant avantageux ne laisse pas le

temps à la recombinaison de casser l'haplotype sur lequel la mutation est apparue. Par conséquent, une signature de la sélection positive pourra être indiquée par un allèle qui a un long et inhabituel déséquilibre de liaison et qui présente une fréquence élevée dans la population (Figure 1.10). Cette signature ne persiste cependant que sur une courte période de temps (< 30 000 ans) du fait du taux de recombinaison élevé observé chez l'homme (Sabeti et al., 2006).

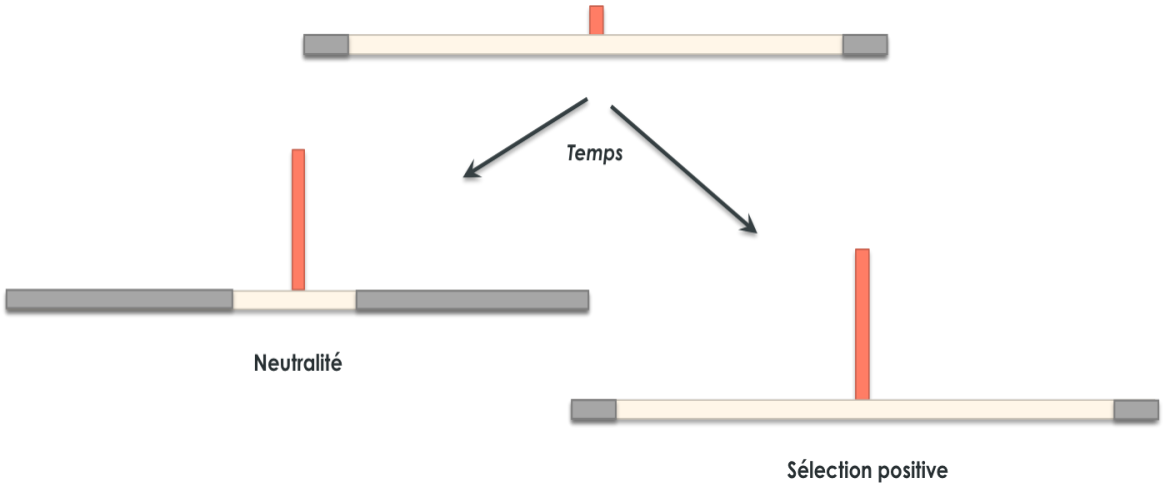


Figure 1.10 | Représentation schématique de l'évolution de la fréquence allélique et de l'étendue du déséquilibre de liaison entourant un variant évoluant sous la neutralité (gauche) ou soumis à une pression de sélection positive de type balayage sélectif (droite). Le chromosome est représenté par la barre grise, l'importance de la fréquence allélique est représentée par la hauteur de la barre rouge et l'étendue du déséquilibre de liaison par la longueur de la barre jaune. Inspiré de Bamshad & Wooding (2003).

Ces dix dernières années, de nombreux tests spécifiquement adaptés à la détection de cette signature moléculaire ont été développés. Ces tests se basent soit strictement sur le niveau de déséquilibre de liaison comme le test LDD (*LD Decay*) (Wang et al., 2006), soit sur la mesure de l'EHH (*Extended Haplotype Homozygosity*), qui est la probabilité que deux chromosomes choisis au hasard dans une population donnée soient homozygotes pour tous les SNPs sur une longue séquence génomique centrée sur le locus sélectionné. Le test princeps introduisant cette dernière statistique est le LRH, pour *long range haplotype* (Sabeti et al., 2002). Par la suite, la statistique EHH a été améliorée à travers un certain nombre de tests apparus de façon

rapprochée : l'iHS (*integrated haplotype score*) qui compare l'EHH entre un allèle ancestral et un allèle dérivé (Voight et al., 2006) ; le WGLRH (*whole genome long range haplotype*) qui détecte les régions où le niveau de déséquilibre de liaison mesuré autour du locus sélectionné est très faible et où les allèles dérivés sont retrouvé à des fréquences élevées (Zhang et al., 2006) ; simultanément l'XP-EHH (Sabeti et al., 2007) et le lnRsb (Tang et al., 2007) qui tous deux contrastent des mesures de l'EHH dans deux populations différentes ; l'EHHST, qui détecte des excès d'homozygotie sur des portions chromosomiques par rapport à ce qui est attendu sous un modèle neutre (Zhong et al., 2010) ; l'XP-EHHST, extension du test précédent pour comparer deux populations (Zhong et al., 2011). Dans cette thèse, nous avons utilisé les deux tests les plus populaires, ayant des capacités complémentaires (Figure 1.11) : l'iHS est puissant pour détecter des balayages sélectifs incomplets, caractérisés par une fréquence de l'allèle sélectionné entre 0,6 et 0,8, et l'XP-EHH des balayages sélectifs complets voire presque complets, dans lesquels la fréquence de l'allèle sélectionné est supérieure à 0,8 (Pickrell et al., 2009; Sabeti et al., 2007).

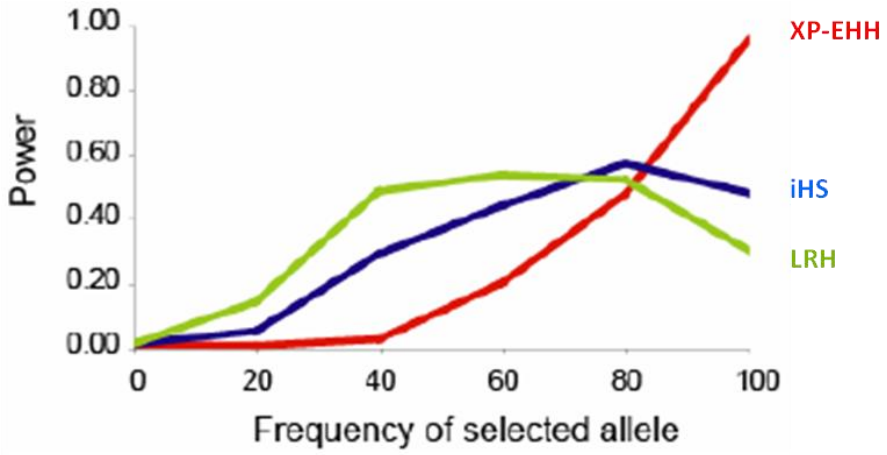


Figure 1. 11 | Puissance de détection d'un balayage sélectif selon la fréquence de l'allèle sélectionné pour les tests de sélection LRH, iHS et XP-EHH. Tiré de (Sabeti et al., 2007).

d. Test basé sur la combinaison de différentes statistiques

Le test CMS, pour *Composite Multiple Signals* a été introduit en 2010 pour essayer de pallier au manque de puissance globale des tests existants pour localiser précisément un signal de sélection sur le génome (Grossman et al., 2010). Il génère un score composite combinant cinq statistiques de tests de sélection positive reposant sur trois signatures distinctes de sélection : les haplotypes étendus (iHS, XP-EHH and ΔiHH), les allèles différenciés (F_{ST}) et les allèles dérivés de fréquence élevée (ΔDAF) [6]. En combinant ainsi des statistiques de tests sensibles à différents types de balayage sélectif (en termes d'âge et de fréquence de l'allèle sélectionné), la statistique CMS permet d'augmenter la capacité de résolution d'un balayage sélectif d'un facteur 100 par rapport à l'usage d'un test individuel (Grossman et al., 2010). Une version adaptée aux données de séquence a également été proposée (Grossman et al., 2013).

3.4 Effets confondants de la démographie

La diversité génétique naturelle des populations interfère directement avec l'interprétation des résultats des tests de sélection naturelle. En effet, les évènements démographiques peuvent laisser des signatures génomiques similaires à celles de la sélection (Przeworski et al., 2000). Par exemple, une réduction importante et rapide de la taille d'une population (*bottleneck*) entraîne une augmentation du nombre d'allèles ayant une fréquence intermédiaire de la même façon que la sélection balancée ; et une expansion de population conduit à une forte réduction de la diversité génétique et un excès d'allèles rares similairement à la sélection positive (Tajima, 1989b).

Il existe cependant une différence dans l'étendue des signatures génomiques. En effet, les phénomènes démographiques affectent tous les marqueurs de la même façon tandis que la sélection naturelle agit en ciblant un locus particulier sur le génome et laisse une signature moléculaire

[6] ΔDAF (*Derived Allele Frequency*) est la différence de fréquence de l'allèle dérivé entre une population donnée et les autres populations incluses dans l'analyse.

détectable sur un sous-ensemble de marqueurs uniquement (Bamshad and Wooding, 2003). Cette caractéristique a été utilisée dans l'approche *outlier* pour distinguer les effets confondants de la démographique sur le génome. Cette approche consiste à comparer ce que l'on observe pour quelques variants d'intérêt par rapport à un grand nombre de marqueurs génétiques répartis sur une partie ou l'ensemble du génome, qui représentent la variation de fond du génome (Kelley et al., 2006). Nous avons adopté cette stratégie et calculé nos statistiques de tests de sélection à l'échelle d'un chromosome entier ou de l'ensemble du génome, obtenant ainsi des distributions empiriques sous H0 (neutralité) pour les différents tests de sélection. Nous avons estimé la significativité statistique des scores obtenus pour des variants d'intérêt en les comparant à ceux obtenus pour l'ensemble des variants de la distribution empirique.

3.5 Mise en pratique de la détection de la sélection naturelle

Deux grandes approches permettent de mettre en évidence des signaux de sélection sur le génome : les études gènes candidats et les études scrutant le génome entier.

Approche gène candidat

Les études gènes candidats considèrent des gènes sur la base de leur fonctionnalité *a priori*. Ces études ont permis de découvrir des signaux d'importance apportant des éclairages significatifs sur l'évolution adaptative de l'homme à des environnements variés. Ainsi ont été détectés des signaux de sélection au niveau du gène *G6PD* conférant une résistance au paludisme en Afrique (Tishkoff et al., 2001) et en Asie du sud-est (Louicharoen et al. 2009) ; du gène *LCT*, impliqué dans le trait de persistance de la lactose en Europe (Bersaglieri et al., 2004) et en Afrique (Tishkoff et al., 2007; Ranciaro et al. 2014) ; des gènes impliqués dans le *pathway HIF* permettant une meilleure oxygénation en haute altitude dans les Andes (Bigham et al., 2009) ; du gène *FOXI1* impliqué dans l'homéostasie évitant la prévention de la déshydratation en Afrique (Moreno-Estrada et al., 2010) ; du cluster *ADH* permettant une protection contre l'alcoolisme en Asie (Han et al., 2007) ; du gène *CYP3A4* impliqué dans le métabolisme de la vitamine D conférant une

résistance augmentée au rachitisme dans les populations non africaines (Schirmer et al., 2006).

La première limite de ces études relève du fait qu'elles nécessitent une hypothèse *a priori* sur la fonction du gène. Pour cela, il est nécessaire d'avoir un minimum de connaissances des relations génotype-phénotype. Or, à l'exception d'un certain nombre de phénotypes relativement bien définis comme la tolérance au lactose, la survie en altitude ou la résistance au paludisme, les maladies affichent une large variabilité phénotypique soutenue par une architecture génétique complexe parfois difficile à appréhender. La définition simple de quelques gènes candidats n'est donc souvent pas évidente. On peut toutefois se concentrer sur certaines grandes catégories de gènes pour lesquelles il est plus ou moins intuitif que la sélection ait joué un rôle majeur, comme les gènes impliqués dans les fonctions immunitaires et la défense de l'organisme (Barreiro and Quintana-Murci, 2010; Fumagalli et al., 2009). C'est ce que nous avons fait dans cette thèse en nous intéressant à l'impact de la sélection naturelle sur les gènes de la réponse aux médicaments.
La seconde limite de ces études est liée aux effets confondants de la démographie, difficiles à discriminer de ceux de la sélection en étudiant une toute petite portion du génome. Toutefois cet obstacle est facilement contournable en adoptant une approche *outlier*, comme nous l'avons fait dans nos travaux.

Recherche des signatures de la sélection naturelle dans le génome entier

Les premiers scans génome entier de la sélection naturelle (*genome wide selection scans*, GWSS) sont apparus dans les années 2000, grâce au développement des méthodes de typage des SNPs à l'échelle pangénomique (Sachidanandam et al., 2001). Ces études se sont concentrées surtout sur la détection des balayages sélectifs incomplets et complets résultant de l'action de la sélection positive sur une unique mutation nouvellement apparue (modèle de type *hard sweep*). Le tout premier GWSS a été réalisé avec la méthode F_{ST} et a permis de détecter de nombreux signaux de sélection concernant notamment des gènes impliqués

dans les fonctions immunitaires et de défense de l'organisme (Akey et al., 2002). Depuis, une grande quantité de GWSS ont été réalisés, en utilisant des tests basés sur : (1) la distorsion du spectre des fréquences alléliques (Carlson et al., 2005; Kelley et al., 2006; Williamson et al., 2007) ; (2) la différenciation génétique des populations (Akey et al., 2002; Barreiro et al., 2008; Chen et al., 2010a) ; (3) la structure haplotypique (Kimura et al., 2007; Lappalainen et al., 2010; Pickrell et al., 2009; Sabeti et al., 2002, 2007; Tang et al., 2007; Teo et al., 2009; Voight et al., 2006; Wang et al., 2006; Zhang et al., 2006; Zhong et al., 2010) ; (4) la combinaison de tests implémentée dans le CMS (Grossman et al., 2013).

Toutes ces études ont permis de détecter un nombre considérable de régions génomiques sous sélection mais souffrent de limitations importantes. Tout d'abord elles sont réalisées en grande majorité sur des données de génotypage dans un nombre limité de populations et n'ont par conséquent accès qu'à une partie de la variabilité génétique des populations. D'autre part, le chevauchement des résultats de ces études est faible et ne peut être expliqué par la seule différence de puissance des méthodes employées en fonction de l'âge des balayages sélectifs détectés. L'une des explications possibles est que chaque étude définit ses propres critères pour déterminer qu'une région du génome est soumise à sélection. Par exemple, il est peu probable qu'une étude qui utilise le test XP-EHH avec un seuil de significativité à 5 % sur des données de génotypage dans quelques populations et une qui utilise le *D* de Tajima avec un seuil de significativité à 1 % sur des données de séquençage dans un grand nombre de populations conduisent à la détection des mêmes régions du génome sous sélection. Ensuite, si ces études fournissent des listes de régions du génome vraisemblablement soumises à sélection, elles pointent peu souvent des gènes, et encore moins des variants génétiques précis. Avec une localisation génomique approximative, il est bien souvent difficile de bien saisir la portée d'un événement sélectif, que ce soit d'un point de vue évolutif (adaptation des populations) ou fonctionnel (rôle dans la détermination de certains phénotypes). Enfin, ces études fournissent une liste de régions candidates pour la sélection mais ne sont quasiment jamais suivies d'analyses d'exploration fonctionnelle des gènes (et surtout des variants candidats) qui

seules permettraient d'améliorer la compréhension des mécanismes à l'origine des maladies génétiques et autres traits phénotypiques d'intérêt.

Nous le verrons dans la partie 2 de cette thèse, l'analyse d'un balayage sélectif aussi bien en termes de répartition géographique que de localisation spatiale, requiert une étude en profondeur sur un grand nombre de populations, des données génétiques denses et plusieurs tests de sélection complémentaires.

3.6 Bases de données utilisées en génétique des populations

Il existe aujourd'hui des bases de données de génotypage (Perlegen, HGDP-CEPH et HapMap) et de séquençage (Projet 1000 Génomes).

Le Projet Perlegen

La base de données Perlegen, créée en 2005, comprend plus de 1,5 million de SNPs génotypés chez 71 individus non apparentés provenant de trois populations d'origine ethnique différente (Américains d'origine africaine et européenne et Chinois Han) (Hinds et al., 2005). L'ensemble des SNPs génotypés ont été identifiés préalablement par un séquençage génome entier de 24 individus représentant ces trois mêmes populations.

Le Panel HGDP-CEPH

Le panel du *Human Genome Diversity Project* (HGDP) se compose de 1 050 individus appartenant à 52 populations réparties dans sept grandes régions géographiques dans le monde (Figure 1.12 et Table 1.1) Il représente la collection la plus complète d'ADN humain au niveau mondial (Cann et al., 2002). Son appellation « Panel HGDP-CEPH » tient au fait que les lignées cellulaires sont conservées au Centre d'Etude du Polymorphisme Humain (CEPH) à Paris. La volonté d'améliorer la compréhension de l'histoire évolutive de l'homme et de la diversité génétique des populations a été à l'origine de la création de ce panel, extrêmement utilisé en génétique des populations humaines. Il a permis d'améliorer la caractérisation de la structure génétique des populations humaines et les causes de cette variabilité (Li et al., 2008b). En 2008, 940 individus non apparentés de ce panel

ont été génotypés sur une puce Illumina d'environ 650 000 SNPs (Illumina HumanHap 650K) (Li et al., 2008b). Nos analyses présentées dans la partie 2 (chapitre 2) de ce document ont utilisé les données de génotypage issues de cette étude.

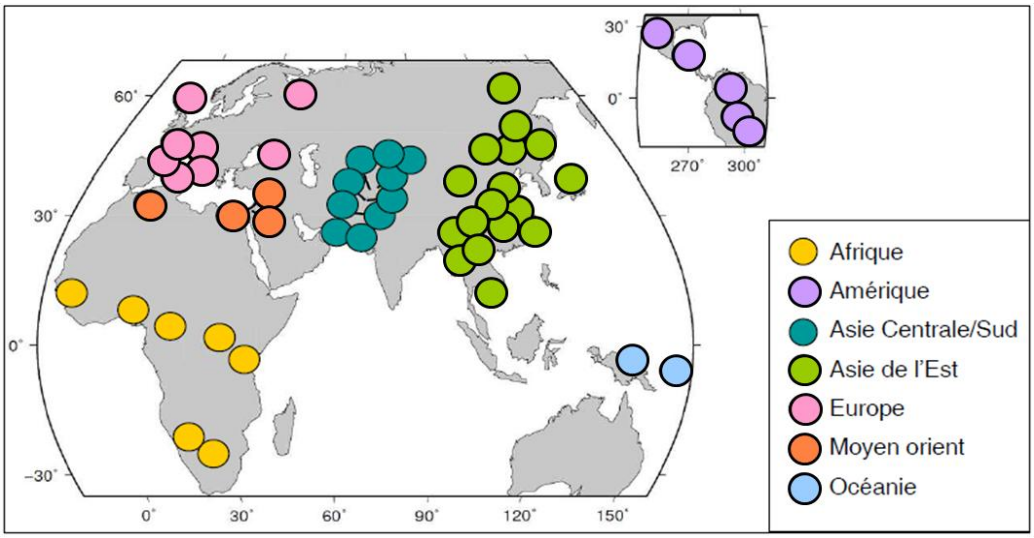


Figure 1.12 | Répartition géographique des 52 populations du panel HGDP-CEPH.

Tableau 1.1| Distribution des 940 individus non apparentés génotypés par Li *et al.* (2008a) dans les 52 populations du panel HGDP-CEPH.

Origine géographique	Population	Nombre d'individus
Afrique		***101***
	San	5
	Pygmées Mbuti	13
	Pygmées Biaka	21
	Yoruba	21
	Mandenka	22
	Bantu du Nord	11
	Bantu du Sud	8
Moyen Orient		***163***
	Mozabites	29
	Bédouins	46
	Palestiniens	46
	Druzes	42
Europe		***157***
	Sardiniens	28
	Toscans	8
	Italiens du Nord	12
	Français	28
	Orcadiens	15
	Basques français	24
	Russes	25
	Adygei	17
Asie Centrale et du Sud		***200***
	Makrani	25
	Balochi	24
	Brahui	25
	Kalash	23
	Burusho	25
	Pathans	22
	Sindhi	24
	Hazara	22
	Ouïghours	10
Asie de l'Est		***228***
	Yakut	25
	Mongola	10
	Tu	10
	Xibo	9
	Oroqen	9
	Hezhen	8
	Daur	9
	Yizu	10
	Naxi	8
	Tujia	10
	Han	44
	She	10
	Miaozu	10
	Dai	10
	Lahu	8
	Japonais	28
	Cambodgiens	10
Océanie		***27***
	Nan Mélanesiens	10
	Papous	17
Amérique		***64***
	Mayas	21
	Pimas	14
	Colombiens	7
	Karitiana	14
	Surui	8

Le Projet HapMap

En 2002 a été lancé le Projet HapMap (*The International HapMap Project*), projet de grande envergure nécessitant une collaboration scientifique internationale ayant pour but de fournir, à travers une base de données publique, un catalogue des variations génétiques communes du génome humain, de déterminer leur distribution et leur fréquence ainsi que leur niveau d'association (les profils de déséquilibre de liaison) dans différentes populations humaines. Plus d'un million de SNPs ont ainsi été génotypés chez 270 individus provenant de quatre populations originaires d'Afrique, d'Europe et d'Asie (International HapMap Consortium, 2003). Aujourd'hui cette base de données fournit l'information concernant plus de 4 millions de SNPs dans

ces quatre mêmes populations (HapMap Phase II, (International HapMap Consortium et al., 2007a)) et plus de 1,6 million de SNPs dans sept populations supplémentaires (HapMap Phase III, (International HapMap 3 Consortium et al., 2010)).

Cette base de données a été largement utilisée par les études d'association pangénomique pour la recherche des gènes et des variants génétiques impliqués dans les maladies complexes ; mais également pour la détection de la sélection naturelle à l'échelle du génome entier (International HapMap Consortium et al., 2007b; Sabeti et al., 2007).

Biais des bases de données de génotypage

Ces trois bases de données ont fait l'objet de quantité d'études permettant un nombre considérable de découvertes importantes pour le monde de la génétique médicale et évolutive. Elles souffrent cependant d'un lourd biais d'échantillonnage des variants (« *ascertainment bias* »). Ce biais est dû à la stratégie choisie pour le choix des SNPs constituant ces bases de données. En effet, les SNPS sont découverts par une première étape de séquençage chez un petit panel d'individus, suivie d'une étape de génotypage ciblé dans un plus grand échantillon, qui ne représente pas forcément le premier échantillon utilisé pour la découverte des variants génétiques. Il en résulte une surreprésentation dans les données finales des variants initialement découverts dans la population d'origine et de fait, une distorsion fréquence-spécifique (Clark et al., 2005). En conséquence, l'estimation des fréquences alléliques, des profils de déséquilibre de liaison et de la différenciation génétique entre populations peuvent être biaisées. Cela est particulièrement vrai dans les populations non représentées dans le panel de découverte des variants.

Le Projet 1000 Génomes (1KG)

Ce projet, d'une ampleur considérable, qui a vu le jour en 2008, avait pour premier objectif de séquencer 1000 génomes humains de différentes origines ethniques. La principale motivation de ce projet était de créer un catalogue complet et détaillé de la variabilité génétique des populations humaines pour, entre autres, favoriser une meilleure compréhension des relations génotype-phénotype. Le but était donc de détecter plus de 95 % de la

variation génétique présente à une fréquence de 1 % dans différentes populations, sur l'ensemble du génome. Les données de séquençage incluent en effet les variants rares et peu fréquents, très peu ou pas représentés dans les données de génotypage. Or ces variants sont très différenciés dans les populations humaines et déterminent de manière importante la structure génétique des populations humaines (Gravel et al., 2011; Mathieson and McVean, 2012). En outre, ces variants sont susceptibles de jouer un rôle majeur dans la prédisposition aux maladies (Marth et al., 2011; Tennessen et al., 2012) et dans la réponse aux médicaments (Nelson et al., 2012).

La phase pilote de ce projet, dont les résultats ont été publiés en 2010, a consisté à générer les données de séquençage à faible profondeur du génome entier de 179 individus appartenant à quatre populations différentes (1000 Genomes Project Consortium et al., 2010). Depuis, la combinaison des données de séquençage peu profond du génome entier (2-6x) et de séquençage profond des exons (50-100x) sont disponibles pour 1 092 individus dont 1 089 non apparentés appartenant à 14 populations de quatre continents (Figure 1.13 et Table 1.2) (1000 Genomes Project Consortium et al., 2012). Ces données représentent plus de 38 millions de variants de type SNV. Ce sont elles que nous avons utilisées dans les travaux présentés dans le chapitre 2 de la partie 2 et dans les parties 3 et 4 de ce manuscrit.

Pour l'étude de la sélection naturelle, les données de séquence possèdent un double avantage par rapport aux données de génotypage :

- L'un est en rapport avec la densité des variants analysés. En effet, les données de séquence permettent un accès à la quasi-totalité de la variabilité génétique, ce qui permet d'explorer les signatures de sélection naturelle à une échelle génétique fine.
- L'autre est celui de ne pas être sensible, par définition, à l'« *ascertainment bias* » qui affecte lourdement les données de génotypage. La présence de ce biais ne permet pas d'appliquer les tests de sélection basés sur le spectre de fréquence allélique dans les données de génotypage comme celles du Panel HGDP. L'utilisation de données

de séquence offre donc la perspective d'une augmentation de la puissance de détection de la sélection.

Nous discuterons plus en détail de ces points dans le chapitre 3 de la partie 2.

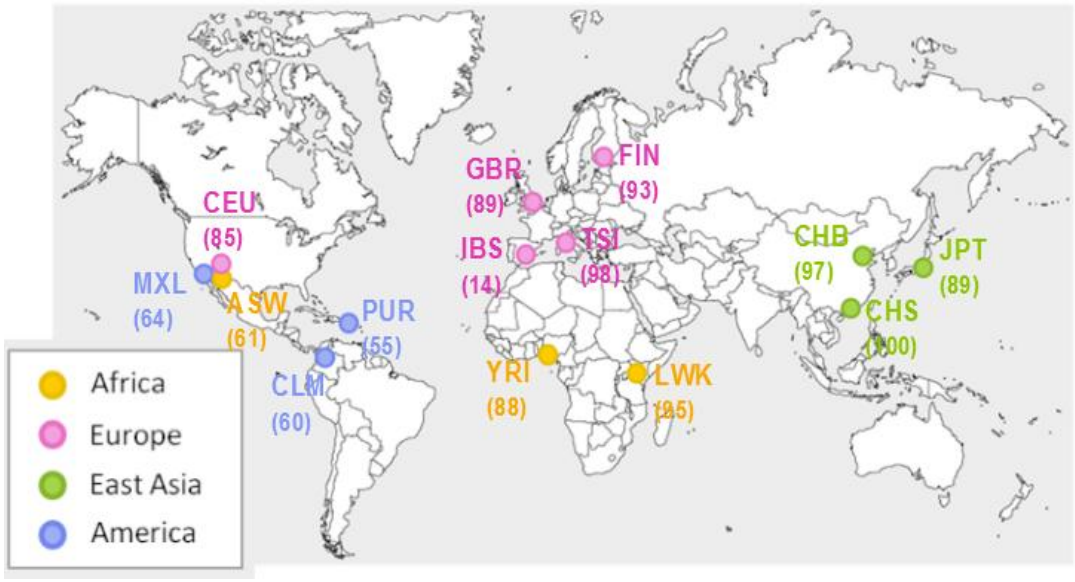


Figure 1.13 | Répartition géographique des 14 populations du Projet 1000 Génomes.

Tableau 1.2 | Distribution des 1089 individus non apparentés dans les 14 populations du Projet 1000 Génomes

Origine géographique	Population	Abbréviation	Nombre d'individus
Afrique			***245***
Nigéria	Yoruba	YRI	88
Kenya	Luhya	LWK	96
Résidents du Sud Ouest des Etats Unis d'origine africaine		ASW	61
Amérique			***179***
Mexique	Mexicains	MXL	64
Puerto Rico	Porto Ricains	PUR	55
Colombie	Colombiens	CLM	60
Asie de l'Est			***286***
Chine	Chinois Han	CHB	97
Japon	Japonais	JPT	89
Chine du Sud	Chinois Han	CHS	100
Europe			***379***
Résidents de l'Utah d'origine d'Europe du Nord et de l'Est		CEU	85
Italie	Toscans	TSI	98
Angleterre et Ecosse	Anglais	GBR	89
Finlande	Finlandais	FIN	93
Espagne	Ibériques	IBS	14

Partie 2

Étude de la différenciation génétique des populations humaines et détection de la sélection positive pour le gène *VKORC1* impliqué dans la réponse aux AVK

Chapitre 1

Généralités sur les anticoagulants oraux de type antivitamine K

Les antivitamines K (AVK) ont été, durant près de 60 ans, les seules molécules anticoagulantes administrables par voie orale. Malgré cet avantage rendant possible des thérapies au long cours – obligatoires dans la plupart des indications des AVK, en particulier en cardiologie – le maniement des AVK est extrêmement délicat. En effet, une marge thérapeutique étroite et une importante variabilité de réponse interindividuelle caractérisent ces molécules. L'équilibre thérapeutique qui en résulte est fragile et long à obtenir. A cette difficulté s'ajoute une iatrogénie élevée et préoccupante, révélée par le grand nombre d'accidents hémorragiques, souvent graves, parfois mortels. De ces aspects du traitement par AVK découle la nécessité d'un suivi biologique laborieux, continu et coûteux. L'introduction récente et progressive des nouveaux anticoagulants oraux, dont le mécanisme d'action ne dépend pas de la vitamine K, n'a pas encore détrôné l'usage massif des AVK dans le traitement des pathologies thrombotiques. Au contraire, l'emploi de ces molécules est en augmentation partout à travers le monde ces dernières années. La découverte cette dernière décennie du rôle prépondérant des facteurs génétiques à l'origine de la variabilité de réponse aux AVK les a placées au premier rang des molécules étudiées en pharmacogénétique.

Nous proposons dans ce chapitre de présenter l'histoire, l'usage et le contexte thérapeutique et pharmacogénétique de ces médicaments, avant d'introduire les raisons qui nous ont conduits à analyser la différenciation génétique des populations humaines et les pressions sélectives sous-jacentes pour le gène *VKORC1*, codant pour la cible pharmacologique directe des AVK.

1. Rappel sur la vitamine K

Historique de sa découverte et caractérisation

La vitamine K a été découverte fortuitement, lors de travaux conduits par le Danois Henrik Dam sur le métabolisme des stérols, et a été immédiatement associée à la coagulation. Observant la survenue d'hémorragies inexpliquées chez des poulets recevant un régime pauvre en cholestérol, Dam fait l'hypothèse de l'existence d'un facteur antihémorragique de type vitamine liposoluble, qu'il propose d'appeler vitamine K en référence au mot danois désignant la *Koagulation* (Dam, 1935; Dam and Schönheyder, 1934).

Il existe trois types de vitamine K, qui diffèrent par leur chaîne latérale (Figure 2.1). La vitamine K1 (phylloquinone), uniquement synthétisée par les plantes, possède une chaîne latérale à 20 atomes de carbone avec une seule double liaison. La vitamine K2 (ménaquinone), synthétisée par les bactéries intestinales, possède une chaîne latérale de 20 à 60 atomes de carbone, dont plusieurs des liaisons sont insaturées. La vitamine K3 (ménadione) n'existe pas à l'état naturel. Il s'agit d'une forme synthétique hydrophile qui ne possède pas de chaîne latérale et qui est convertie en vitamine K2 dans le corps (Odyakov et al., 1992). En 1939, la vitamine K1 est isolée dans la luzerne par Dam et ses associés ; la vitamine K2 dans la farine de poisson par le groupe d'Edward Doisy (McKee et al., 1939a). La synthèse chimique de la vitamine K1 est réalisée indépendamment par trois groupes différents en 1939 (Almquist and Klose, 1939; Fieser, 1939; McKee et al., 1939b), celle de la vitamine K2 l'année suivante par le groupe d'Edward Doisy (Binkley et al., 1940). Ces travaux d'importance sur la vitamine K ont valu à Dam et Doisy d'être récompensés par le Prix Nobel de Médecine en 1943.

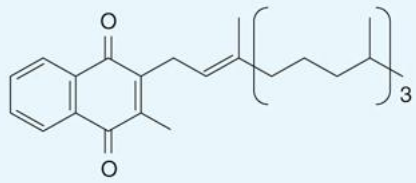
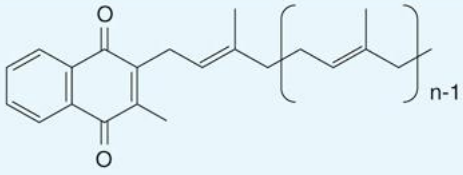
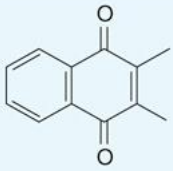
	Formule chimique
Phylloquinone : vitamine K_1	
Ménaquinone : vitamine K_2	
Ménadione : vitamine K_3	

Figure 2.1 | Formule chimique des différentes formes de la vitamine K, phylloquinone ou vitamine K_1, ménaquinone ou vitamine K_2 et ménadione ou vitamine K_3. Tiré de (Moreau et al., 2012).

Rôle physiologique

Les vitamines K (K1, K2 et K3) forment un groupe de vitamines liposolubles, servant de co-facteurs pour l'étape de gamma-carboxylation, modification post-traductionnelle indispensable à l'activité biologique des protéines dites vitamine K-dépendantes.

La liste des protéines vitamines K-dépendantes, n'a cessé de s'accroître ces dernières années, du fait de la découverte de nouvelles fonctions de ces molécules, parfois vitales. On en connaît aujourd'hui 17, dont sept sont impliquées dans la coagulation. Néanmoins le rôle physiologique de la plupart d'entre elles n'est pas encore bien compris (Beulens et al., 2013). Les processus physiologiques concernés par la vitamine K, en plus de la coagulation sanguine, sont nombreux : citons par exemple le remodelage osseux, les processus de calcification de l'os, des vaisseaux sanguins et des tissus conjonctifs, les évènements de signalisation cellulaire (adhésion, croissance, division, prolifération, migration, apoptose, spécialisation), le contrôle de la réponse inflammatoire, la résistance à l'insuline (Berkner and Runge, 2004; Bügel, 2008; Ferland, 2012; Vermeer et al., 2004).

Apport en vitamine K

L'apport en vitamine K est principalement endogène : elle est synthétisée par les bactéries de la flore intestinale à partir de la fermentation des aliments. La vitamine K est également apportée par l'alimentation : la vitamine K1 est essentiellement retrouvée dans les légumes verts (chou, laitue, brocolis, épinards, asperges) ainsi que les huiles végétales (olive, soja,…) (Booth and Suttie, 1998) et la vitamine K2 dans les fromages fermentés, le soja (natto), la viande et les œufs (Elder et al., 2006; Tsukamoto et al., 2000).

2. Histoire des AVK

Découverte et caractérisation

Les antivitamines K (AVK) ont été découverts à la suite d'intoxications dévastatrices du bétail au Canada et dans le Nord des États Unis au début des années 20 (Mueller and Scheidt, 1994; Wardrop and Keeling, 2008). Les animaux, nourris avec du foin contenant du trèfle doux (*melilotus alba* et *Melilotus officinalis*) mourraient d'hémorragies spontanées causées par la molécule 3,3'-méthylène-bis[4-hydroxycoumarine] (un antivitamine K), provenant de l'oxydation, sous l'effet de l'humidité, de la coumarine présente à l'état naturel dans la plante. Depuis cet épisode dramatique, cette maladie est connue sous le nom de *sweet clover disease* (Roderick, 1929; Schofield, 1924).

Il a fallu attendre les travaux de l'américain Karl Link des années 30 à 40 pour que l'isolation de cette molécule anticoagulante, le 3,3'-méthylène-bis[4-hydroxycoumarine], soit obtenue (Huebner and LINK, 1941; LINK, 1959). La molécule est baptisée alors le dicoumarol.

Premières utilisations

En 1945, Link a eu l'idée d'utiliser les dérivés coumariniques comme raticides. A cet effet, le dicoumarol étant trop lent à agir, des molécules semblables plus efficaces sont développées. La warfarine, particulièrement active, a vu le jour en 1948, et est utilisée en tant que raticide en provoquant des hémorragies intestinales aiguës fatales (LINK, 1959). Par rapport au dicoumarol, la warfarine a l'avantage d'être une molécule ayant une grande solubilité dans l'eau et une biodisponibilité orale

importante. Le nom *warfarin* est dérivé de WARF (*Wisconsin Alumni Research Foundation*) qui finançait les travaux de Link, et *–arin*, de *coumarin*.
Elle est utilisée chez l'homme avec succès pour la première fois en 1950 en tant que médicament anticoagulant. Sa commercialisation sous le nom de Coumadin® date de 1954, et la première étude clinique sur la warfarine de 1955 (POLLOCK, 1955). Pour l'anecdote, cette même année, le président des États-Unis Dwight Eisenhower en reçoit suite à un infarctus du myocarde.
Aujourd'hui, la warfarine est devenue la référence des traitements au long terme des pathologies thrombotiques artérielles et veineuses.

Nomenclature

Il existe actuellement une dizaine de molécules antivitamine K de synthèse, regroupées au sein de deux familles (Tableau 2.1) : la famille des coumarines, à laquelle appartiennent la warfarine (Coumadine®) et l'acénocoumarol (Sintrom®), et les AVK non-coumariniques, dont la fluindione (Préviscan®). Cette dernière molécule correspond à plus de 80 % des prescriptions d'AVK en France ; alors que l'utilisation de la warfarine prédomine aux États-Unis et en Grande-Bretagne et celle d'acénocoumarol dans le reste de l'Europe. Aujourd'hui la warfarine reste la molécule de référence à laquelle sont consacrées la majorité les études disponibles sur les AVK.

Tableau 2.1 | AVK commercialisés en France.

Famille pharmacologique	Dénomination commune internationale	Nom commercial
Coumariniques	Acénocoumarol	Sintrom® 4mg Minisintrom® 1mg
	Warfarine	Coumadine® 2mg Coumadine® 5mg
Dérivé de l'indanedione	Fluindione	Préviscan® 20mg

3. Mécanisme d'action des AVK

Le mécanisme d'action des AVK n'a été élucidé qu'en 1974, à la suite de travaux décrivant la carboxylation des facteurs de la coagulation vitamine K-dépendants (Stenflo et al., 1974). En 1978, deux équipes ont, de manière indépendante, découverts que la warfarine agissait par inhibition de la *vitamine K époxide réductase* (VKORC1) (Bell, 1978; Whitlon et al., 1978).

3.1 Pharmacodynamie des AVK

Comme l'illustre la Figure 2.2, les AVK exercent leur effet anticoagulant de manière indirecte en inhibant l'action de l'enzyme VKORC1 qui transforme la vitamine K oxydée en vitamine réduite. Cette vitamine K réduite est un co-facteur de la *gamma-glutamyle carboxylase* (GGCX), enzyme impliquée dans la gamma-carboxylation hépatique des facteurs de la coagulation vitamine K-dépendants (II, VII, IX et X), et de trois inhibiteurs (protéines C, S et Z). En l'absence de gamma-carboxylation, les facteurs de la coagulation sont inactifs.

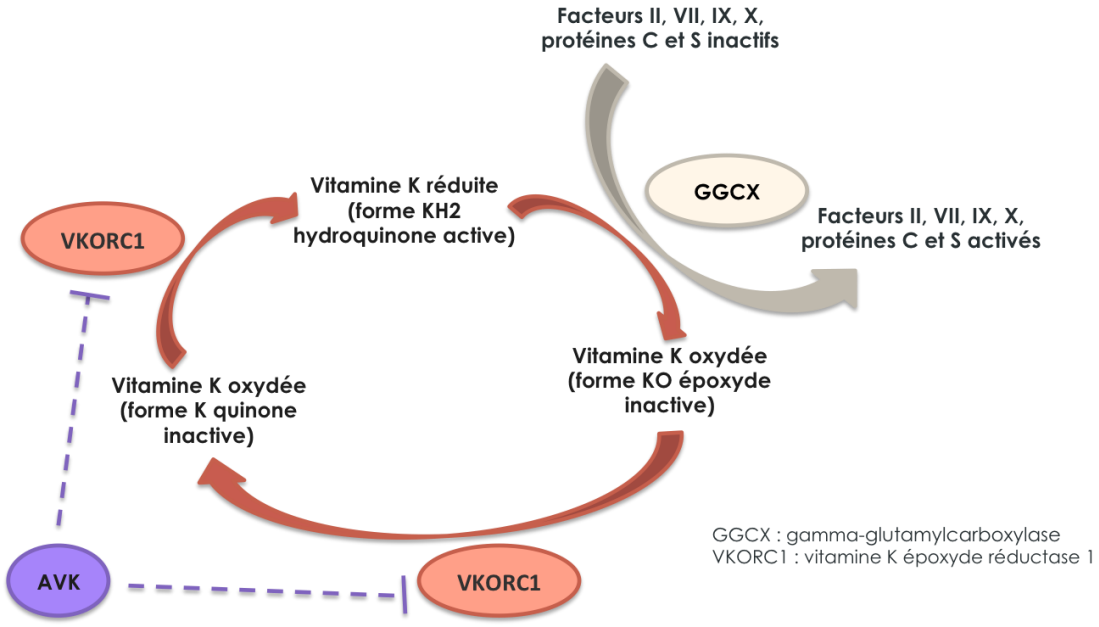


Figure 2.2 | Représentation schématique du mécanisme d'action des AVK.

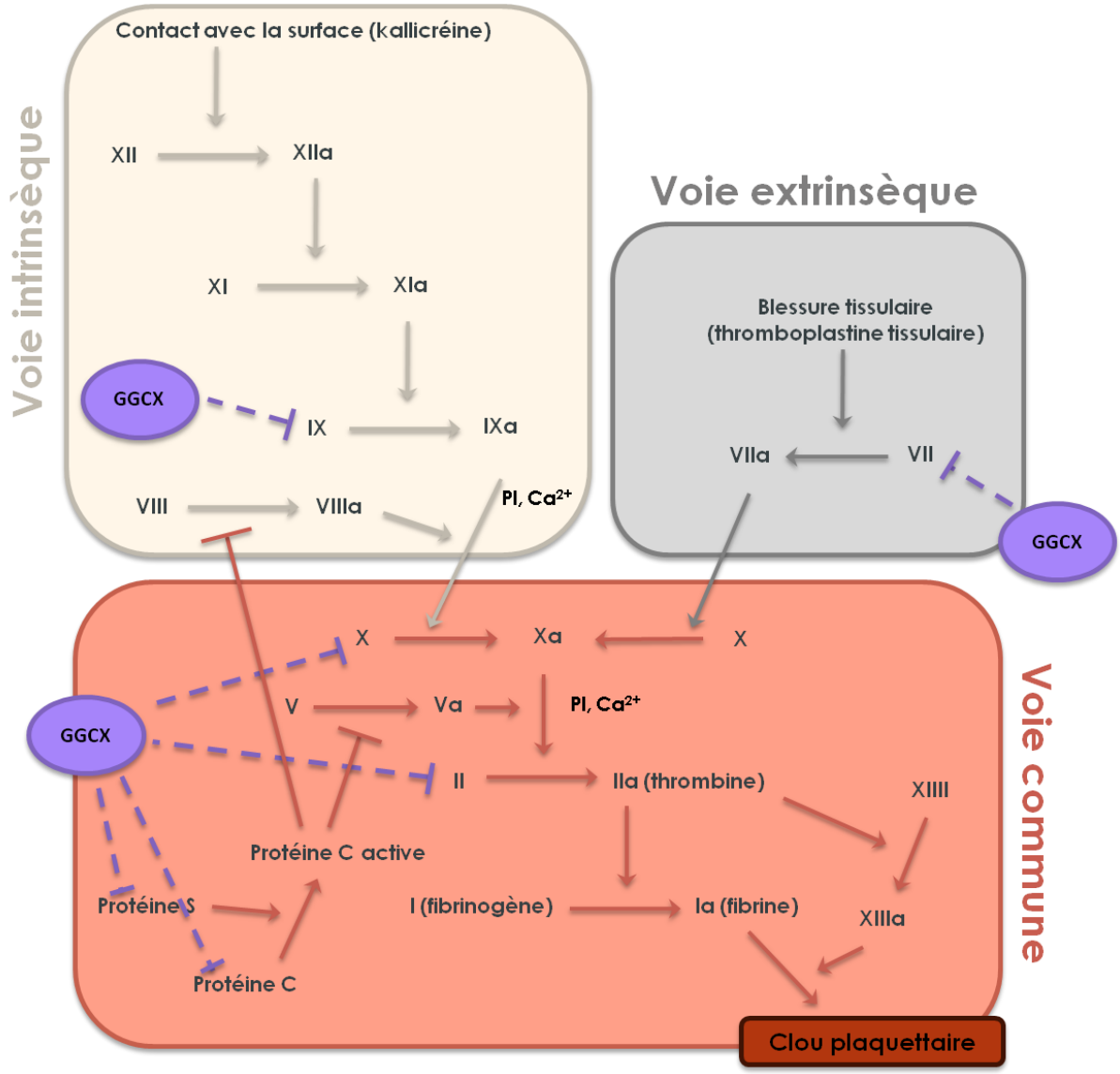


Figure 2.3 | Représentation schématique de la cascade de la coagulation.
PL : phospholipide plaquettaire.

3.2 Pharmacocinétique des AVK

L'absorption digestive des AVK est importante et rapide et s'effectue principalement au niveau de l'estomac et du jéjunum. La concentration plasmatique maximale est atteinte en 90 minutes. Les AVK se lient de façon très importante à l'albumine dans le plasma (92 % à 98 %). Seule la fraction libre est active et subit un métabolisme oxydatif hépatique important via les enzymes du cytochrome P450 : CYP2C9 principalement, et CYP3A4, CYP1A1, CYP1A2 et CYP2C19 dans une moindre mesure (Daly and King, 2003). L'élimination se fait soit par la bile sous forme de métabolite inactif, soit par voie urinaire sous forme de produit pur lié à l'albumine.

4. Indications thérapeutiques et usage des AVK

Dès leur découverte, les anticoagulants oraux de type AVK ont constitué les molécules de choix dans la prévention et le traitement des troubles thrombo-emboliques (DOUGLAS, 1955). La preuve de la nécessité d'un traitement anticoagulant après une embolie pulmonaire a été apportée en 1960 (BARRITT and JORDAN, 1960). Ils permettent en particulier, en empêchant la formation d'un thrombus, de réduire très efficacement l'incidence des accidents vasculaires cérébraux (AVC) chez les patients souffrant de fibrillation auriculaire, le plus fréquent des troubles du rythme cardiaque. Ce dernier affecte jusqu'à 1 % de la population générale, et est responsable de 15 % des AVC chez le sujet jeune (Burnett et al., 2013; Hylek et al., 2003) (HAS, 2007).

4.1 Indications thérapeutiques des AVK

Selon l'Agence Nationale de Sécurité du Médicament et des produits de santé (ANSM, 2012), les AVK sont utilisés dans :

- la prévention et le traitement de la thrombose veineuse profonde et de l'embolie pulmonaire, en relais à l'héparine ;
- la prévention des complications thromboemboliques en rapport avec certains troubles du rythme auriculaire (fibrillations auriculaires, flutter, tachycardie atriale), certaines valvulopathies mitrales, les prothèses valvulaires ;
- la prévention des complications thromboemboliques des infarctus du myocarde compliqués : thrombus mural, dysfonction ventriculaire gauche sévère, dyskinésie emboligène..., en relais à l'héparine.

Environ 80 % des prescriptions des AVK s'inscrivent dans le cadre de traitements au long cours, mais il existe des traitements de courte durée (3 à 6 mois), qui concernent principalement la prévention et le traitement des thrombophlébites et des embolies pulmonaires.

4.2 Utilisation des AVK

La consommation des AVK est colossale et en augmentation constante : en France elle a presque doublé entre 2000 et 2010, passant de 7,6 à 13,8 millions de boîtes vendues.

En 2011 en France, 1,8 % de la population couverte par le régime général a perçu au moins un remboursement d'AVK. En extrapolant ces données à la population française, le nombre de sujets traités en 2011 est estimé à environ 1,1 million.

Aux États-Unis, il est estimé que plus de 1 % de la population générale est sous AVK et plus de 8 % des personnes âgées de plus de 80 ans (Pirmohamed, 2006).

5. Variabilité de réponse aux AVK

Du fait d'une marge thérapeutique étroite et d'une grande variabilité de réponse interindividuelle, les AVK sont des molécules extrêmement délicates à utiliser. Il n'est pas possible notamment de définir à l'avance la posologie optimale à administrer au patient. Pour la warfarine par exemple, les doses nécessaires pour obtenir un même degré d'anticoagulation varient considérablement entre les individus : bien que la dose moyenne de warfarine pour atteindre l'effet thérapeutique recherché soit de 4 à 5 mg par jour, ces doses peuvent varier de 0,5 mg à plus de 20 mg par jour en fonction des patients (Wadelius et al., 2005).

5.1 Surveillance biologique du traitement par AVK

Afin de vérifier que le patient ne sort pas de la zone thérapeutique, il est absolument indispensable de réaliser une surveillance biologique, systématique et minutieuse, du traitement par AVK. Ce suivi a pour objectif de rechercher la dose optimale de la molécule permettant de maintenir l'efficacité de l'anticoagulation tout en limitant le risque hémorragique.

La surveillance du traitement par AVK s'effectue par la mesure de l'INR (*international normalized ratio*), examen de laboratoire réalisé à partir d'un prélèvement de sang. Il est adopté par l'Organisation Mondiale de la Santé (OMS) en 1983 (OMS, 1983).

L'INR mesure le temps de coagulation du patient et le compare à celui d'un sujet qui ne reçoit pas de traitement AVK. Il est calculé comme suit :

$$INR = \left(\frac{TQpatient}{TQtémoin}\right)^{ISI}$$; avec :

- $TQ_{patient}$: le temps de Quick[7] mesuré pour le plasma du patient à tester ;
- $TQ_{témoin}$: le temps de Quick témoin, qui correspond à la moyenne géométrique des temps de Quick mesurés sur le plasma frais d'au moins 30 sujets adultes sains ne recevant pas de traitement antivitamine K ;
- ISI = l'indice de sensibilité international spécifique du réactif thromboplastine utilisé.

L'INR n'a pas d'unité. Du fait de la normalisation par l'indice ISI, sa valeur ne dépend pas du réactif utilisé ; par conséquent plusieurs mesures pour un même sujet, bien que réalisées dans des laboratoires différents, peuvent être comparées entre elles sans problème. Cependant il est conseillé de toujours faire mesurer son INR dans le même laboratoire.

L'INR cible varie en fonction de l'indication du traitement et de l'âge du patient. La zone thérapeutique se situe habituellement entre 2 et 3 (INR cible de 2,5) pour la plupart des indications, mais peut être plus élevée chez les patients porteurs de prothèses mécaniques. En dehors de maladies du foie, l'INR d'un sujet non traité par AVK est inférieur ou égal à 1,2.

Instauration d'un traitement par AVK

La dose initiale, toujours probatoire, doit être aussi proche que possible de la dose d'équilibre (une fois l'effet thérapeutique atteint et le traitement stabilisé), qui dépend de la molécule employée.

Le premier contrôle de l'INR s'effectue après la 3e prise d'AVK. Il permet de dépister les réactions d'hypersensibilité individuelle : un INR supérieur à 2 annonce un surdosage et nécessite une adaptation posologique immédiate. Le second contrôle, réalisé 3 à 6 jours après le premier selon les cas, permet d'apprécier l'efficacité de l'anticoagulation.

[7] Le temps de Quick est le temps de coagulation du plasma, exprimé en secondes par rapport au temps témoin (mesuré sur le plasma d'environ trente sujets normaux)

Adaptation du traitement par AVK

En début de traitement, l'INR est mesuré tous les 2 à 4 jours, jusqu'à ce que l'INR ait atteint la valeur cible souhaitée. L'ajustement de la posologie des AVK s'effectue par paliers, en contrôlant l'INR tous les 2 à 4 jours jusqu'à stabilisation de sa valeur sur deux contrôles successifs. Tant que l'INR cible n'est pas atteint, la posologie d'AVK est ajustée. Elle est maintenue une fois l'INR cible atteint et stabilisé. Les contrôles de l'INR sont alors progressivement espacés pour atteindre un contrôle tous les 3 à 4 semaines une fois le traitement anticoagulant équilibré.

5.2 Iatrogénie des AVK

En dépit de leur utilisation massive dans le monde et du suivi biologique par la mesure de l'INR, l'emploi des AVK n'est pas sans risque. Le principal risque associé à un traitement anticoagulant, quel qu'il soit, est le risque hémorragique, qui apparaît en cas de surdosage du médicament. Cet effet secondaire est inhérent au mode d'action des anticoagulants qui ralentissent la coagulation sanguine. Il peut s'agir d'hémorragies non graves (hématome, épistaxis, gingivorragie) ou graves (hémorragie ou hématome intracérébral, hématome du psoas, hémorragie intra-abdominale, hémorragie intra-articulaire). Une méta-analyse de 33 études a estimé que des hémorragies graves ou fatales arrivent dans respectivement 7,2 et 1,3 cas par 100 patients-années de traitement par warfarine aux États-Unis (Pirmohamed, 2006).

Le sous-dosage des AVK expose au risque de thrombose, à l'origine de la prescription.

En raison de leurs complications hémorragiques et de leur difficulté d'emploi, le potentiel iatrogène des AVK est extrêmement élevé. En effet, ces médicaments sont responsables d'un nombre très élevé d'accidents, qui les classent aujourd'hui au premier rang des molécules iatrogènes dans le monde. Cette constatation représente un problème de santé publique majeur, ayant de multiples répercussions en termes de morbi-mortalité, de coût, et de prévention.

Accidents iatrogéniques

Les enquêtes nationales sur les événements indésirables graves associés aux soins (ENEIS2) engagées par la DREES (Direction de la recherche, des études, de

l'évaluation et des statistiques) démontrent que les AVK arrivent en France au premier rang des médicaments responsables d'accidents iatrogènes graves. Globalement, on estime que 5 000 à 6 000 décès par an en France sont causés par la survenue d'une hémorragie sous AVK (ANSM, 2012).

Aux États-Unis, une étude menée en 2007 montre que 17,3 % des consultations d'urgence liées à des effets indésirables des traitements médicamenteux de patients de plus de 65 ans, sont dues à la warfarine, soit plus de 30 000 visites par an (Budnitz et al., 2007). En Inde, une étude a montré récemment que les anticoagulants étaient les médicaments responsables de 15,1 % des effets indésirables conduisant à une hospitalisation (Haile et al., 2013).

Impact économique

Il est difficile de trouver des données précises sur les conséquences économiques des effets indésirables imputables exclusivement aux AVK. Néanmoins, comme nous l'avons expliqué dans la première partie de cette thèse, il est désormais reconnu que les coûts engendrés par la iatrogénie médicamenteuse de manière générale sont extrêmement élevés.

Prévention des accidents iatrogéniques liés aux AVK

Le nombre très élevé d'accidents, pour la plupart évitables, a incité les autorités de santé à mettre en œuvre différentes initiatives au niveau national pour tenter d'améliorer cet état de fait. La prévention des risques iatrogènes constitue à l'heure actuelle un enjeu majeur de sécurité sanitaire.

Depuis 1998, en France, l'ANSM s'est engagée dans un programme d'évaluation, de prévention et de gestion des risques iatrogènes médicamenteux évitables, qui a conduit notamment à des publications sur le bon usage des AVK à destination des professionnels de santé (ANSM, 2012). En 2008, la Haute Autorité de Santé (HAS) a publié des recommandations professionnelles sur la prise en charge des surdosages, des situations à risque et des accidents hémorragiques chez les patients traités par AVK (Burnett et al., 2013).

Les anticoagulants font également l'objet d'un suivi au niveau européen dans le cadre des Plans de Gestion des Risques (PGR).

Par ailleurs, pour aider à la prise en charge et à la surveillance du traitement par AVK, un carnet d'information et de suivi « *Vous et votre traitement anticoagulant par AVK* » destiné au patient a été élaboré en 2004 (AFSSAPS, 2008). Généralement fourni par le médecin prescripteur, son utilisation est recommandée dans l'AMM des spécialités concernées. Il contient des informations pour le patient lui rappelant les règles de bon usage des AVK, et lui permet de noter régulièrement toutes les informations pertinentes en rapport avec le traitement (prises, les résultats de l'INR,…), permettant d'améliorer le suivi du traitement. En outre la délivrance des AVK s'accompagne d'une éducation thérapeutique du patient à certains points clés du traitement par AVK.

5.3 Facteurs influençant la variabilité de réponse aux AVK

La grande variabilité interindividuelle de l'effet anticoagulant des AVK, mesurée par la dose d'AVK nécessaire pour obtenir l'INR cible, est expliquée par différents types de facteurs.

Facteurs non génétiques

Il s'agit des facteurs physiopathologiques usuels, tels que l'âge, le sexe, le poids, la présence de pathologies associées (insuffisance hépatique, rénale, …) (Kamali et al., 2004; Sconce et al., 2005; Wynne et al., 1995).

Pour un même degré d'anticoagulation, les posologies d'AVK requises sont plus faibles chez les sujets âgés que chez les sujets jeunes (Debray et al., 2003; Redwood et al., 1991). Il est important de considérer ce facteur afin d'ajuster la posologie en fonction de l'âge du patient : en effet, il a été démontré qu'entre 20 et 90 ans, la dose optimale de warfarine diminue de 8 à 17 % par décennie (Gage et al., 2004; Kamali et al., 2004; Sconce et al., 2005; Wynne et al., 1995).

L'influence du sexe sur la réponse aux AVK n'est pas claire. Alors que certaines études observent une augmentation de la fréquence des saignements chez les femmes traitées par warfarine, d'autres ne détectent pas de différence homme-femme (Levine et al., 2004). Un certain nombre de pathologies ont été associées avec la survenue de saignements au cours d'un traitement par warfarine, incluant l'hypertension artérielle, les pathologies cardiovasculaires, les accidents vasculaires cérébraux, l'insuffisance rénale et les cancers (Levine et al., 2004).

Par ailleurs, les AVK sont associés à un risque élevé d'interactions médicamenteuses, avec un grand nombre de molécules telles que l'acide acétylsalicylique, les anti-inflammatoires non stéroïdiens pyrazolés, des inducteurs enzymatiques du cytochrome P450 (alcool chronique, tabac, millepertuis) ou des inhibiteurs enzymatiques (miconazole) (Hirsh et al., 2003; Holbrook et al., 2005; Wells et al., 1994).

L'observance du patient à son traitement est un facteur affectant la stabilité du traitement par AVK (van der Meer et al., 1997). Il a été montré que la non-adhérence au traitement ainsi que le non respect des recommandations alimentaires représentent les causes majeures (36 %) des INR en dehors des valeurs thérapeutiques (Waterman et al., 2004). Parmi les facteurs associés au niveau d'adhérence ont été identifiés l'âge, le fait d'habiter à proximité d'un laboratoire d'analyse médicale, d'être fumeur, le nombre d'hospitalisations ainsi que le type et le stade de la maladie (Pamboukian et al., 2008; Waterman et al., 2004). Étant donné la dangerosité des AVK, l'éducation thérapeutique du patient au bon usage des AVK est fondamentale pour garantir, non seulement des thérapies efficaces, mais aussi une diminution du risque de survenue d'effets indésirables.

Facteurs génétiques

Ensemble, les facteurs physiopathologiques et environnementaux ne permettent d'expliquer qu'une faible partie de la variabilité interindividuelle de la réponse aux AVK (17 à 22 %) (Gage et al., 2008). Des polymorphismes dans les gènes codant pour les protéines impliquées aux différents niveaux de cette réponse (transporteurs, enzymes du métabolisme, récepteur, etc.) peuvent moduler le niveau et la qualité de la réponse individuelle à un traitement par AVK. Ils sont discutés de manière détaillée ci-après.

6. Facteurs génétiques

On sait aujourd'hui que les facteurs génétiques constituent les déterminants majeurs de la grande variation de réponse aux AVK. Si une part de cette variabilité demeure encore inexpliquée aujourd'hui, leur découverte a néanmoins permis d'améliorer considérablement la précision de la posologie initiale d'AVK à administrer à chaque patient et par voie de conséquence, l'efficacité et la sûreté de ces traitements.

6.1 Études gènes candidats

Les premières études recherchant les facteurs génétiques de la réponse aux AVK ont utilisé une approche gène candidat et se sont tout naturellement tournées vers les candidats les plus évidents : les gènes *CYP2C9* et *VKORC1*.

CYP2C9

Le gène *CYP2C9* joue un rôle dans la réponse aux AVK par un mécanisme pharmacocinétique : il code en effet pour l'enzyme CYP2C9 du cytochrome P540, majoritairement responsable du métabolisme hépatique des AVK. Les deux principaux polymorphismes de *CYP2C9* impliqués dans la variation de la réponse aux AVK sont les allèles *CYP2C9*2* et *CYP2C9*3*, respectivement définis par des variations aux locus rs1799853 et rs1057910, qui exhibent seulement 12 % et 5 % respectivement de l'activité enzymatique normale (c'est-à-dire celle correspondant à l'allèle de référence *CYP2C9*1*), conduisant à une diminution de la clairance des AVK (D'Andrea et al., 2008; Lee et al., 2002). En conséquence, les individus porteurs de ces allèles requièrent des doses d'AVK réduites par rapport aux individus porteurs de l'allèle *CYP2C9*1*. Ensemble, ces variants permettent d'expliquer environ 12 % (5-22 % selon les études) de la variabilité de réponse aux AVK (D'Andrea et al., 2008; Wadelius et al., 2005).

VKORC1

Le gène *VKORC1* joue quant à lui un rôle par un mécanisme pharmacodynamique. En effet, ce gène code pour la sous-unité 1 du complexe *vitamine K époxyde réductase* (VKOR1), cible pharmacologique des AVK (Li et al., 2004; Rost et al., 2004). Il a été découvert en 2004 par clonage positionnel comme étant le gène impliqué dans le déficit combiné en facteurs de la coagulation vitamine K dépendants et dans la résistance à la warfarine.

Initialement, le premier polymorphisme de *VKORC1* associé à la variabilité interindividuelle de la réponse aux AVK était le SNP intronique rs9934438 (1173C>T) (D'Andrea et al., 2005). Peu de temps après, il a été montré qu'il existait un déséquilibre de liaison complet entre les allèles du SNP rs9934438 et ceux d'un autre SNP rs99239231 (-1639G>A), situé dans le promoteur de *VKORC1*, dans le site de liaison E-box d'un facteur de transcription (Yuan et al., 2005). Par ailleurs, une étude haplotypique de *VKORC1* a révélé plusieurs autres polymorphismes en fort déséquilibre de liaison avec -1639G>A et 1173C>T, et associés au taux d'ARNm et à la dose requise d'AVK (Rieder et al., 2005) Se posait alors la question de savoir le(les)quel(s) de ces polymorphismes étaient responsables des différences biologiques dans l'activité de l'enzyme VKORC1. Étant donné l'emplacement du variant rs9923231 dans un site de liaison d'un facteur de transcription, il était probable qu'il pouvait entrainer une modification de l'activité de promoteur. Deux études distinctes ont testé cette hypothèse avec des essais de type gène luciférase rapporteur, mais ont conduit à des résultats contradictoires, car cette technique ne permet pas d'évaluer de manière consistante la régulation et l'expression des gènes, laissant ouverte la question de savoir si ce polymorphisme était fonctionnel ou non (Bodin et al., 2005; Yuan et al., 2005). Une exploration en profondeur des mécanismes moléculaires sous-tendant l'activité de VKORC1 est parvenue à identifier le polymorphisme fonctionnel entre les deux principaux candidats (rs9923231 et rs9934438), en utilisant la technique d'immunoprécipitation de la chromatine, révélant l'allèle dérivé -1639A comme étant celui régulant l'expression de l'ARNm du gène *VKORC1*, en inhibant l'activité du promoteur dans le foie par l'intermédiaire d'une modification de la chromatine (Wang et al., 2008). A lui seul, ce variant explique environ 28 % (6-37 % selon les études) de la variation de dose de warfarine ou d'acénocoumarol (Bodin et al., 2005; D'Andrea et al., 2005; Gage et al., 2008; Lee et al., 2009; Sconce et al., 2005).

La Figure 2.4 présente les doses quotidiennes de warfarine (A) et hebdomadaires d'acénocoumarol (B) requises en fonction des génotypes des variants fonctionnels de *VKORC1* (rs9923231 pour la warfarine ou rs9934438 pour l'acénocoumarol) et de *CYP2C9* (rs1799853 et rs1057910). On peut remarquer que quelque soit le génotype de *CYP2C9*, les individus homozygotes pour l'allèle dérivé A du variant rs9923231 de *VKORC1* (ou l'allèle T du variant rs9934438) requièrent des doses plus faibles d'AVK

que les individus homozygotes pour l'allèle ancestral, et que les individus hétérozygotes pour ce variant requièrent des doses intermédiaires.

Par ailleurs, la dose requise d'AVK est plus faible chez les patients porteurs des deux allèles mutés *CYP2C9*2* (rs1799853) et *CYP2C9*3* (rs1057910) que chez les patients homozygotes sauvages (*1/*1), et ce quelque soit le génotype de *VKORC1*. Cependant, il semble pour la réponse à l'acénocoumarol que l'impact du polymorphisme de *CYP2C9* soit plus important lorsque le variant de *VKORC1* est à l'état homozygote pour l'allèle ancestral. De plus, la réduction de la dose d'AVK est plus importante chez les individus portant l'allèle *CYP2C9*3* que l'allèle *CYP2C9*2*, ce qui peut s'expliquer par le fait que ce dernier a un effet plus atténué sur la réduction de la clairance des AVK que l'allèle *CYPC9*3* (Lee et al., 2002).

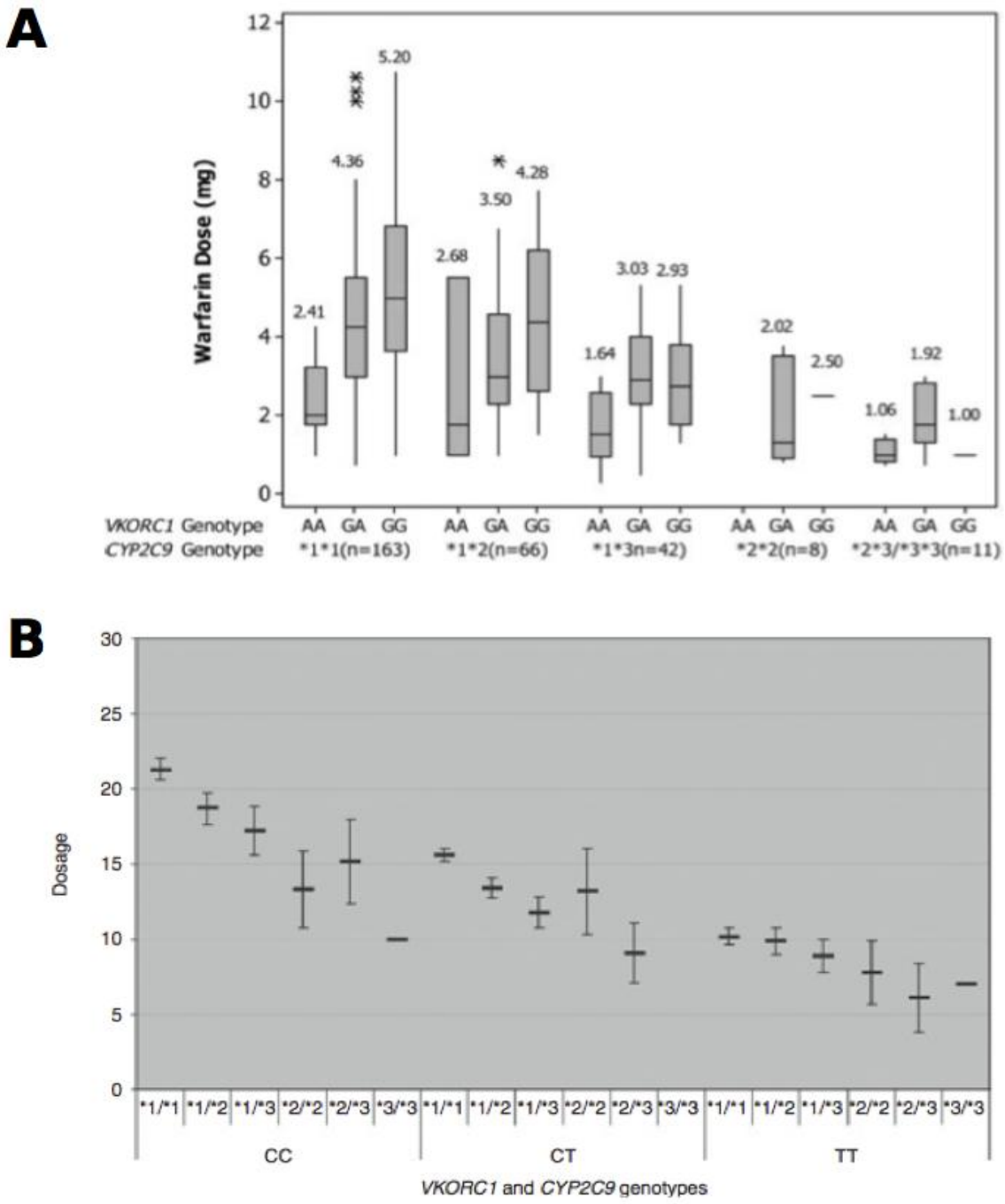


Figure 2.4 | Boîtes à moustache montrant la distribution de la dose de molécule AVK selon le génotype du variant fonctionnel -1639G>A (A) ou 1173C>T (B) de VKORC1 et des allèles CYP2C9*2 et *3. (A) Dose quotidienne de warfarine. Les nombres au-dessus des rectangles indiquent les valeurs moyennes de dose de warfarine. Les lignes verticales indiquent les valeurs minimales et maximales, les astérisques représentent les valeurs extrêmes. Tiré de (Sconce et al., 2005). **(B) Dose hebdomadaire d'acénocoumarol.** Les lignes verticales indiquent les valeurs minimales et maximales. Tiré de (Teichert et al., 2009b).

6.2 Études d'association génome entier

Entre 2008 et 2010, quatre études d'association pangénomiques, ont été réalisées dans le but d'identifier de nouvelles cibles impliquées dans la réponse à la warfarine et à l'acénocoumarol dans des populations d'origine européenne ou japonaises (Cha et al., 2010; Cooper et al., 2008; Takeuchi et al., 2009; Teichert et al., 2009). Elles ont confirmé le rôle majeur des gènes *CYP2C9* et *VKORC1* et mis en évidence celui, bien plus faible, de deux autres gènes impliqués dans le métabolisme hépatique des AVK. Ainsi, prendre en compte le génotype porté au SNP rs2108622 de *CYP4F2* permet d'expliquer environ 1 % de la variation de dose de warfarine et d'acénocoumarol, et celui au SNP rs1998591 de *CYP2C18* environ 1 % de la variation de réponse de l'acénocoumarol uniquement. L'association spécifique du variant de *CYP2C18* avec la dose d'acénocoumarol s'explique par le fait que l'enzyme CYP2C18 intervient dans le métabolisme de l'acénocoumarol et non dans celui de la warfarine. Ensemble, les cinq polymorphismes fonctionnels des gènes *VKORC1*, *CYP2C9*, *CYP4F2* et *CYP2C18* mentionnés (Figure 2.5) peuvent expliquer près de 50 % de la variation de dose des AVK, en plus des facteurs cliniques classiques.

Cependant ces variables génétiques ne permettent pas d'expliquer qu'une faible part de la variabilité de la dose de warfarine dans les populations d'origine africaine par rapport aux populations d'origine européenne et asiatique (Limdi et al., 2010). Une étude a montré que le variant de *VKORC1* explique 18 % (30 % en ajoutant les deux variants de *CYP2C9*) de la variabilité de la réponse aux AVK dans les populations d'origine européenne, alors qu'il en explique seulement 5 % (8 % avec *CYP2C9*) dans les populations d'origine africaine (Limdi et al., 2008b).

Afin d'identifier les facteurs génétiques déterminant de façon plus spécifique la dose de warfarine dans ces populations, une étude d'association pangénomique récente, réalisée sur des individus africains-américains, a détecté une association avec le variant rs12777823, situé sur le chromosome 10 au niveau du locus CYP2C, indépendant des polymorphismes connus de *CYP2C9* (Perera et al., 2013). Cette association semble être spécifique des individus d'origine africaine, car elle n'est pas retrouvée pour des individus d'origine européenne, japonaise, ou égyptienne. De même, l'association avec le variant de *CYP4F2* n'est pas retrouvée dans cette population. L'inclusion du variant rs12777823 dans l'algorithme développé par

l'IWPC, présenté ci-après, permet d'augmenter la puissance de prédiction de la dose de warfarine de 5 % chez ces individus (Perera et al., 2013).

Le Tableau 2.2 récapitule les différentes études GWAS sur les déterminants génétiques de la dose des AVK. On voit bien que dans la majorité des cas, elles concernent la warfarine, AVK de référence international.

6.3 Algorithmes pharmacogénétiques de prédiction des doses de warfarine et acénocoumarol

Les tous premiers algorithmes pharmacogénétiques de prédiction de la dose de warfarine incluaient uniquement les variants du gène *CYP2C9* et étaient capables de n'expliquer qu'une part modérée de la variabilité de réponse à cette molécule (< 40 %) (Tableau 2.3). Les algorithmes incluant le gène *VKORC1* sont apparus en 2005, pour la réponse à la warfarine. Plus puissants, ces algorithmes permettaient de prédire, en tenant compte du génotype de ces deux gènes et de variables cliniques plus de 50 % de la variance de la dose requise de warfarine. Tous ces algorithmes étaient jusqu'alors établis dans des populations d'origine ethnique particulière. En 2009, l'IWPC a mis au point un algorithme pharmacogénétique de prédiction de la dose initiale de warfarine, basé sur les génotypes des trois variants fonctionnels de *VKORC1* et *CYP2C9* (Klein et al., 2009), qui permet d'estimer la dose initiale de warfarine de manière plus robuste qu'un algorithme basé uniquement sur les données cliniques, quelle que soit l'origine ethnique des patients (Limdi et al., 2010).
Depuis, plusieurs algorithmes pharmacogénétiques intégrant cette information génétique en plus de variables cliniques et démographiques ont vu le jour, dans des populations de différentes origines ethniques. En moyenne, ces algorithmes sont capables d'expliquer approximativement entre 50 et 60 % de la variabilité de la dose d'AVK (Borobia et al., 2012; Gage et al., 2008; Langley et al., 2009; Liu et al., 2012b; Shin and Cao, 2011).
Cependant, bien que l'information génotypique soit significativement associée avec la prédiction de la dose thérapeutique de warfarine à tous les stades du traitement, il semble que leur contribution diminue au fur et à mesure des semaines de thérapie. En effet une étude a mis en évidence qu'à J0, les trois variants fonctionnels de *VKORC1* et *CYP2C9* permettent à eux seuls d'expliquer 43 % de la

variation de dose, mais que cette proportion est réduite à 12 % au septième jour de traitement, puis 3,9 au 14e et enfin 1,4 % au 21e (Ferder et al., 2010). Après plusieurs semaines de traitement, la dose thérapeutique de warfarine est mieux prédite par l'historique des doses du traitement que par les facteurs génétiques.

Un site internet www.warfarindosing.org à destination des médecins et cliniciens a été créé pour estimer la dose journalière de warfarine à administrer sur la base de variables cliniques et du statut génétique du patient. Deux algorithmes pharmacogénétiques renvoyant des recommandations de doses très similaires y sont implémentés, dont celui du IWPC (Gage et al., 2008; Klein et al., 2009).

Par ailleurs, la modification du label de la warfarine, conseillant de tenir compte de l'information pharmacogénétique de *CYP2C9* et *VKORC1* a été approuvée aux États-Unis par la FDA en 2010. En 2011, le CPIC (*Clinical Pharmacogenetics Implementation Consortium*) du NIH a publié des recommandations à destination des cliniciens ayant pour objectif de les encourager à utiliser un test de génotypage de *CYP2C9* et *VKORC1* et de les assister dans l'interprétation de l'information génétique, pour estimer la dose de warfarine à administrer au patient garantissant un INR stable à 2-3 (Johnson et al., 2011).
Concernant la réponse à l'acénocoumarol, le premier algorithme pharmacogénétique a été développé en 2008. A l'heure actuelle, seulement cinq algorithmes ont été développés pour cette molécule, dont quatre dans des populations européennes, alors qu'il en existe 32 pour la warfarine développés dans des populations du monde entier, illustrant la plus forte concentration des efforts de recherche sur cette dernière molécule (Tableau 2.3).

Tableau 2.2 | Résumé des études d'association génome-entier réalisées sur la réponse aux AVK.

Auteurs	Date	Population d'étude	Effectifs des échantillons d'étude	Molécule AVK	Variants signifcativement associés	% de variance totale expliquée par ces variants
Cooper	2008	Origine caucasienne	Découverte : 181 Réplication : 379	Warfarine	- rs10871454 (en DL complet avec rs9923231 dans *VKORC1*) - rs4917639 dans *CYP2C9* (tag les allèles CYP2C9*2 et *3)	- rs10871454 : 25 % - rs4917639 : 9 %
Teichert	2009	Origine caucasienne	Découverte : 1451 Réplication : 287	Acénocoumarol	- rs10871454 (en DL complet avec rs9923231 dans *VKORC1*) - rs4086116 dans *CYP2C9* (en DL complet avec rs4917639)) - rs2108622 (*CYP4F2*) - rs1998581 (*CYP2C18*)	- rs10871454 : 32 % - rs4086116 : 6,9 % - rs2108622 : 1,1 % - rs1998581: 0,9 %
Takeuchi	2009	Suédois	Découverte : 1053 Réplication : 588	Warfarine	- rs9923231 (*VKORC1*) - rs1799853 et rs1057910 (*CYP2C9*2 et *3) - rs2108622 (*CYP4F2*)	- rs9923231 : 28,3 % - rs1799853 : 3,8 % - rs1057910 : 8% - rs2108622 : 1,1 %
Cha	2010	Japonais	Découverte :1515 Réplication : 444	Warfarine	- rs9923231 (*VKORC1*) - rs10509680 (*CYP2C9*) - rs2108622 (*CYP4F2*)	- rs9923231 : 25,7 % - rs10509680 : 1,8 % - rs2108622 :1,4 %
Perera	2013	Africains-Américains	Découverte : 533 Réplication : 43)	Warfarine	rs12777823 (en amont de *CYP2C18*)	- rs12777823 : 5 %

Reference	Country	*n*	Type	Genetic parameters	Clinical parameters	R^2	MAE (mg day^{-1})
Warfarin							
Gage *et al.* 2004 [79]	USA	369	M	CYP2C9	Age, gender, BSA, race, target INR, CM	39%	–
Hillman *et al.* 2004 [78]	USA	453	M	CYP2C9	Age, BSA, valve replacement, diabetes	34%	–
Kamali *et al.* 2004 [80]	UK	121	M	CYP2C9	Age	20%	–
Sconce *et al.* 2005 [81]	UK	297	M	CYP2C9, VKORC1	Age, height	54%	–
Carlquist *et al.* 2006 [82]	USA	213	M	CYP2C9, VKORC1	Age, gender, weight	45%	–
Herman *et al.* 2006 [90]	Slovenia	165	M	CYP2C9, VKORC1	Age, BSA	60%	–
Takahashi *et al.* 2006 [92]	Japan	365	M	CYP2C9, VKORC1	Age, weight	57%	–
Tham *et al.* 2006 [91]	Singapore	107	M	CYP2C9, VKORC1	Age, weight	60%	–
Gage *et al.* 2008 [83]	USA	1015	M	CYP2C9, VKORC1	Age, BSA, race, target INR, CM, smoking	57%	1.3
Perini *et al.* 2008 [104]	Brazil	390	M	CYP2C9, VKORC1	Age, weight, heart valve prosthesis, thromboembolic disease, CM	50%	0.99
Wu *et al.* 2008 [85]	USA	92	M	CYP2C9, VKORC1	Age, gender, weight, height, race, CM, smoking	59%	–
IWPC 2009 [84]	Various	4043	M	CYP2C9, VKORC1	Age, height, weight, race, CM	47%	1.19
Huang *et al.* 2009 [97]	China	266	M	CYP2C9, VKORC1	Age, BSA	45%	–
Sasaki *et al.* 2009 [93]	Japan	45	M*	CYP2C9, VKORC1	*	94%*	–
Wadelius *et al.* 2009 [40]	Sweden	1496	M	CYP2C9, VKORC1	Age, gender, race, CM	59%	–
Harada *et al.* 2010 [94]	Japan	97	M	CYP2C9, VKORC1, CYP4F2	Age, white blood cell count, CM	49%	–
Lenzini *et al.* 2010 [107]	Various	969	R	CYP2C9, VKORC1	Age, BSA, race, stroke, target INR, diabetes, CM, dose and INR values	60%	0.79
Wells *et al.* 2010 [87]	Canada	249	M	CYP2C9, VKORC1, CYP4F2	Age, BMI, height, exercise level, CM	58%	–
Avery *et al.* 2011 [106]	UK	671	I	CYP2C9, VKORC1	Age, height, weight, CM	42%	–
Cho *et al.* 2011 [95]	Korea	130	M	CYP2C9, VKORC1	Age, BSA, CM	60%	–
Choi *et al.* 2011 [96]	Korea	564	M	CYP2C9, VKORC1, CYP4F2, GGCX	Age, BSA, gender, INR	35%	–
Gong *et al.* 2011 [86]	UK and Canada	167	I and M	CYP2C9, VKORC1, CYP4F2	Age, weight, gender, CM	42%	1.49
Suriapranata *et al.* 2011 [101]	Indonesia	85	M	CYP2C9, VKORC1	Age, weight, height	21%	–
You *et al.* 2011 [98]	China	100	M	CYP2C9, VKORC1	Age, weight, vitamin K intake	68%	–
Zambon *et al.* 2011 [89]	Italy	274	M	CYP2C9, VKORC1, CYP4F2	Age, BSA	65%	0.97
Cini *et al.* 2012 [88]	Italy	55	M	CYP2C9, VKORC1	Age, height, weight, gender, smoking, vegetable intake, indication, diabetes	44%	1.42
Horne *et al.* 2012 [108]	Various	2022	R	CYP2C9, VKORC1	Age, BSA, CM, stroke, target INR, dose and INR values	72%	0.71
Pathare *et al.* 2012 [103]	Oman	212	M	CYP2C9, VKORC1	Age, weight, gender, indication	62%	0.26
Pavani *et al.* 2012 [102]	India	240	M	CYP2C9, VKORC1	Age, BMI, gender, vitamin K intake	89%	–
Ramos *et al.* 2012 [105]	Puerto Rico	163	M	CYP2C9, VKORC1	Age, indication, CM, dose-adjusted INR	67%	0.79
Wei *et al.* 2012 [99]	China	325	M	CYP2C9, VKORC1, CYP4F2	Age, weight, previous thromboembolism, CM	52%	–
Xu *et al.* 2012 [100]	China	207	R	CYP2C9, VKORC1, CYP4F2	Age, BSA, target INR and INR values	54%	0.59
Acenocoumarol							
Markatos *et al.* 2008 [61]	Greece	98	M	CYP2C9, VKORC1	Age, gender, CM	55%	–
Van Schie *et al.* 2011 [27]	The Netherlands	375	I and M	CYP2C9, VKORC1	Age, height, weight, gender, CM	53%	0.52
Borobia *et al.* 2012 [111]	Spain	147	M	CYP2C9, VKORC1, CYP4F2, APOE	Age, BMI, CM	61%	0.52
Rathore *et al.* 2012 [110]	India	125	M	CYP2C9, VKORC1, CYP4F2, GGCX	Age, weight, height, BSA, gender smoking, indication	41%	0.71
Cerezo-Manchado *et al.* 2013 [112]	Spain	973	M	CYP2C9, VKORC1, CYP4F2	Age, BSA, gender	50%	–

Tableau 2.3 | Algorithmes pharmacogénétiques de prédiction de la dose de warfarine et acénocoumarol publiés. Tiré de (Verhoef et al., 2014). **PKPD model. M, Maintenance dose; R, Refinement; I, Initiation dose; CM, concomitant medication; MAE, mean absolute error.*

6.4 Variabilité inter-populationnelle dans la réponse aux AVK

D'importantes différences dans la dose requise d'AVK sont observées entre les populations humaines. En effet, le IWPC reporte des doses hebdomadaires de warfarine moyennes de 21 mg pour les individus asiatiques, de 31,5 mg pour les européens, et de 40 mg pour les individus d'origine africaine (Limdi et al., 2010). Par ailleurs, les algorithmes pharmacogénétiques tenant compte des génotypes de *CYP2C9* et de *VKORC1* sont connus pour être moins performants pour prédire la dose requise de warfarine chez les individus d'origine africaine que chez ceux d'origine européenne et asiatique. Ces différences peuvent s'expliquer par l'hétérogénéité de distribution des variants fonctionnels de ces gènes entre les populations humaines (Schelleman et al., 2008).

Au cours de mon travail de Master 2 en 2010, nous avons génotypé les cinq principaux variants fonctionnels des gènes *VKORC1*, *CYP2C9, CYP4F2* et *CYP2C18* dans les 52 populations du Panel HGDP-CEPH afin d'analyser leur distribution dans les populations humaines. La Figure 2.5 présente ces distributions.

Elles révèlent notamment que la fréquence de l'allèle dérivé A du variant fonctionnel rs9923231 -1639G>A de *VKORC1*, qui confère une sensibilité augmentée aux AVK, varie beaucoup entre les populations humaines. A l'exception des San, la fréquence de cet allèle de sensibilité est très faible (3 %) en Afrique, alors qu'elle atteint pratiquement une valeur maximale de 1 dans les 17 populations d'Asie de l'Est. Dans les autres régions géographiques, cet allèle est retrouvé à des fréquences intermédiaires. La différence observée entre les San et les autres populations africaines peut s'expliquer soit par le faible effectif de l'échantillon étudié qui ne permet pas d'avoir une estimation précise et fiable des fréquences alléliques dans cette population, soit par un effet fort de la dérive génétique dû au faible effectif efficace de cette population.

Les autres variants de sensibilité aux AVK étudiés sont retrouvés à des fréquences plus homogènes. L'allèle dérivé du variant rs1799853 (définissant l'allèle *CYP2C9*2*) a une fréquence mondiale faible (5 %). Il n'est quasiment présent qu'au Moyen-Orient, en Europe et en Asie Centrale et du Sud. La fréquence de l'allèle dérivé C du SNP rs1057910 (définissant l'allèle *CYP2C9*3*) est également faible: elle ne dépasse pas 20 % dans aucune des 52 populations échantillonnées. Sa répartition, en revanche,

est plus large que l'allèle *CYP2C9*2* : il est observé dans tous les continents excepté l'Afrique.

La distribution des variants rs1998591 de *CYP2C18* et rs2108622 de *CYP4F2* est très homogène au sein des 52 populations du Panel. La fréquence de l'allèle dérivé C du SNP rs1998591 de *CYP2C18* est élevée à l'exception des deux populations d'Océanie. Celle de l'allèle dérivé T de *CYP4F2* est modérée. Cet allèle est peu présent en Afrique et en Amérique.

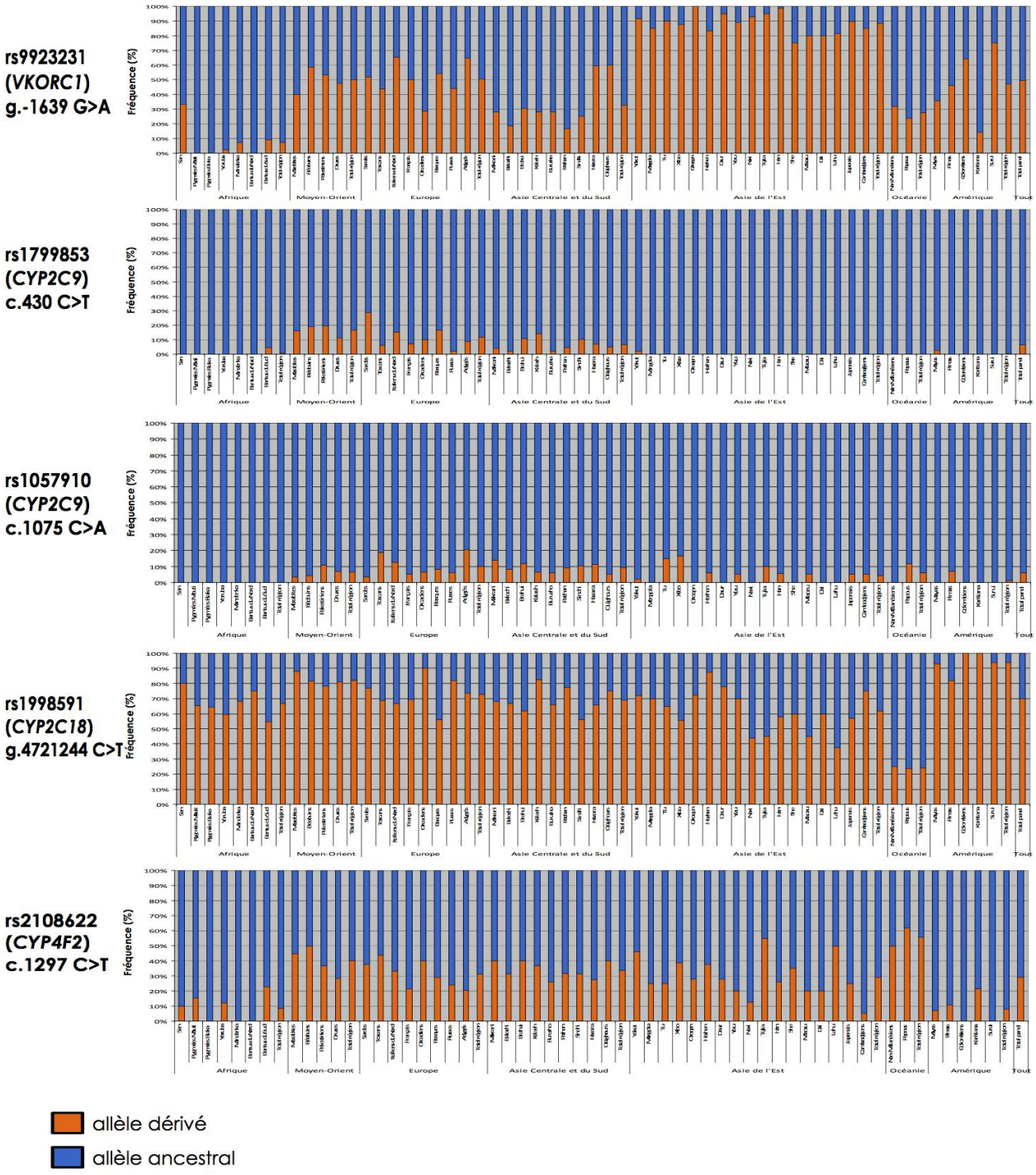


Figure 2.5 | Distribution des cinq variants fonctionnels associés à la réponse aux AVK au sein des 52 populations humaines du panel HGDP-CEPH. Les allèles dérivé et ancestral sont représentés en orange et bleu, respectivement. Tiré de Patillon B (2010) « *Diversité génétique des populations humaines et variabilité de réponse aux AVK* », Mémoire de Master 2.

Chapitre 2

Analyse haplotypique mondiale de *VKORC1* et détection de la sélection positive

En 2010, dans leur article paru dans la revue *Journal of Human Genetics*, Ross *et al.*, ont étudié la distribution mondiale des variants majeurs connus pour jouer un rôle dans la réponse à la warfarine : rs9923231 dans *VKORC1*, rs1799853 et rs1057910 dans *CYP2C9*, et rs2108622 dans *CYP4F2* en génotypant également ces variants dans les 52 populations du panel HGDP-CEPH (Ross et al., 2010), et ont confirmé les distributions alléliques présentées dans le chapitre précédent.

Face au profil de distribution particulièrement hétérogène du variant fonctionnel rs9923231 de *VKORC1*, évocateur de la sélection naturelle, Ross *et al.* (Ross et al., 2010) ont alors recherché des signatures de sélection naturelle dans les populations asiatiques (CHB+JPT) de HapMap pour lesquelles ils pouvaient disposer de données de génotypage denses dans cette région génomique. Ils ont mis en œuvre trois tests de détection de la sélection positive : le test *locus-specific branch length* (LSLB), mesurant le niveau de différenciation génétique entre populations ; le test *log of the ratio of heterozygosities* (Ln*RH*), comparant le degré de diversité génétique entre deux populations ; et la statistique *D* de Tajima, basée sur le spectre de fréquence allélique. En se basant sur la distribution génome entier de ces trois statistiques, Ross *et al.* ont détecté pour chacun des tests, des signaux significatifs de sélection au niveau du gène *VKORC1*.

Bien qu'étant la première à mettre en évidence de la sélection positive sur ce gène, l'étude de Ross *et al.* souffre de quelques limitations ne lui permettant pas de correctement caractériser l'événement sélectif détecté en Asie de l'Est. Tout d'abord, puisqu'ils ont testé la sélection à partir des données HapMap, ils n'ont vu un signal que dans les deux populations asiatiques chinoise (CHB) et japonaise (JPT) et ils n'ont pas pu étudier la répartition géographique mondiale de l'évènement sélectif – en particulier dans les régions proches de l'Asie. Ensuite, cette étude ne décrit pas non plus la localisation génomique exacte de la signature de sélection, du fait notamment des tests de sélection employés qui ne sont pas adaptés à localiser précisément un signal de sélection. Par conséquent cela ne permet pas d'éliminer la possibilité que *VKORC1* ne soit pas la cible directe de la sélection, mais soit en déséquilibre de liaison avec un gène voisin qui l'aurait entrainé à des fréquences élevées en Asie de l'Est par le phénomène d'autostop génétique. Par ailleurs, le *D* de Tajima employé par Ross *et al.* n'est pas parfaitement adapté pour détecter de la sélection positive sur des données de génotypage : il est en effet sensible au biais de détermination des SNPs (*ascertainement biais*) qui affecte lourdement les données de génotypage (*cf.* partie 1).

Récemment, des tests de sélection spécifiquement adaptés pour détecter des signatures génomiques de la sélection positive sur des données de génotypage ont été développés (Chen et al., 2010a; Sabeti et al., 2002, 2007; Voight et al., 2006). Ces tests sont plus appropriés pour identifier de la sélection positive récente que les tests appliqués par Ross *et al.*, en particulier s'il s'agit d'un évènement de sélection de type '*hard sweep*', c'est-à-dire d'un balayage sélectif presque complet, comme cela semble être le cas pour *VKORC1*, au regard des fréquences de l'allèle dérivé A du variant rs9923231 en Asie de l'Est.

Dans l'étude présentée ci-après, nous avons exploré en profondeur les signatures génomiques laissées par la sélection naturelle au locus de *VKORC1*, en employant des tests de sélection spécifiquement adaptés pour

détecter de la sélection positive sur des données de génotypage pangénomique.

1. Résumé de l'article 1

L'étude présentée ci-après a pour objectifs :

(1) de fournir une description de la diversité haplotypique mondiale du gène *VKORC1* à partir de sept variants génotypés dans le Panel HGDP-CEPH, afin notamment d'identifier le(s) haplotype(s) sur lequel(s) ségrégent le polymorphisme fonctionnel rs9923231 et dans quelles populations.

(2) de déterminer précisément le rôle joué par la sélection naturelle dans l'évolution du gène *VKORC1* chez l'Homme, en décrivant notamment :

   a. la répartition géographique précise de l'événement sélectif identifié par Ross *et al.* en Asie de l'Est,
   b. l'étendue de la région génomique autour du locus *VKORC1* affectée par la sélection naturelle
   c. les gènes dans cette région génomique le plus probablement ciblés par la pression de sélection.

Pour réaliser cette étude, nous avons utilisé les données de génotypage du panel HGDP-CEPH, comprenant 52 populations dans le monde, dont 17 échantillonnées en Asie de l'Est (*cf.* partie 1, chapitre 2). Les 940 individus non apparentés de ce panel ont été génotypés en 2008 par Li *et al.* (2008) sur une puce Illumina 650K, qui comprend un variant de *VKORC1* (rs7294) (Li et al., 2008c). Nous avons génotypé six variants supplémentaires de *VKORC1* chez ces individus, disposant ainsi d'un total de sept variants dans *VKORC1* pour procéder à la reconstruction des haplotypes.

Les haplotypes ont été reconstruits avec le programme fastPHASE : parmi les sept haplotypes retrouvés à une fréquence supérieure à 1 % dans au moins une des sept grandes régions géographiques, un seul porte l'allèle dérivé A du variant fonctionnel rs9923231 -1639G>A, qui confère une sensibilité

augmentée aux AVK. Cet haplotype, le plus fréquent au niveau global (49,7 %), est retrouvé à des fréquences particulièrement élevées en Asie de l'Est (89,6 %), très faibles en Afrique Sub-saharienne (4,4 %), et à des fréquences intermédiaires dans les autres régions géographiques (de 27,8 à 51,2 %).

La construction du réseau d'haplotypes à ce locus a révélé que cet haplotype différait des autres par la présence de deux polymorphismes : le variant fonctionnel rs9923231, et le variant intronique rs9934438, qui est en déséquilibre de liaison complet avec rs9923231 dans toutes les régions géographiques.

L'étude approfondie de la sélection positive a requis la mise en œuvre de plusieurs tests de sélection, spécialement conçus pour détecter de la sélection positive. Les tests F_{ST} et XP-CLR sont basés sur la mesure de la différenciation génétique des populations ; les tests iHS et XP-EHH sur le partage d'haplotypes identiques entre les populations sur de grandes régions génomiques. Chacun de ces tests possède une capacité différente à identifier une signature génomique de sélection positive selon la période écoulée depuis la pression de sélection et la fréquence à laquelle l'allèle avantagé est parvenu sous l'effet de la sélection. Par conséquent, ces différents tests sont complémentaires, et leur application simultanée nous a permis d'augmenter la puissance et la précision d'identification d'un signal de sélection.

L'analyse combinée des différents signaux obtenus pour chaque test dans une région étendue de 2 Mb centrée sur le gène *VKORC1* dans les sept régions géographiques nous a permis de délimiter le signal de sélection à une région de 505 kb en fort déséquilibre de liaison, dans laquelle sont situés 25 gènes dont *VKORC1*. L'évènement sélectif détecté qui aurait débuté il y a approximativement 4 500 ans semble concerner l'ensemble des populations d'Asie de l'Est et uniquement cette région du monde. Il s'agit effectivement d'un balayage sélectif presque complet qui a affecté cette région du génome, événement assez rare durant l'histoire adaptative de l'homme.

Une exploration approfondie des scores les plus extrêmes obtenus dans cette région nous a permis d'affiner la localisation spatiale du signal de sélection à une région de 45 kb comprenant quatre gènes : *BCKDK*, *MYST1* (*KAT8*), *VKORC1* et *PRSS8*. Il n'est pas possible de déterminer, du fait du déséquilibre de liaison très fort qui existe dans cette région du génome lequel de ces gènes a été la cible de la sélection naturelle.

Il est possible que *VKORC1* soit la cible directe de la sélection, auquel cas l'avantage sélectif pourrait être en rapport avec le métabolisme de la vitamine K, ou bien avec l'exposition à une molécule de type AVK présente à l'état naturel dans l'environnement chimique des populations d'Asie de l'Est.

Cependant, nous ne pouvons pas non plus exclure la possibilité qu'un polymorphisme situé dans un autre gène avoisinant *VKORC1* ait été la cible de la sélection. Il aurait alors entraîné par un phénomène d'autostop génétique le variant fonctionnel de *VKORC1*. Cette hypothèse est d'autant plus probable que parmi les quatre gènes, on en trouve deux impliqués dans la réponse immunitaire (*MYST1* et *PRSS8*) qui sont des bons gènes candidats à l'action de la sélection naturelle (Barreiro and Quintana-Murci, 2010). Cependant aucun de ces deux gènes ne contenait de variant fonctionnel à fréquence élevée en Asie de l'Est dans les données HapMap.

Bien qu'il ne soit pas possible, du fait du déséquilibre de liaison important dans la région sélectionnée, de déterminer avec exactitude si *VKORC1* représente la véritable cible de la sélection, la pression sélective qui a agi dans cette région du génome en Asie de l'Est il y a environ 4 500 ans est responsable de la distribution très hétérogène au sein des populations humaines du variant fonctionnel du promoteur qui confère la sensibilité augmentée aux AVK. Aujourd'hui, cela se traduit par des différences significatives dans les doses requises d'AVK entre les populations humaines. Cette étude illustre bien l'importance du rôle de la sélection naturelle dans la détermination de la variabilité inter-populationnelle des phénotypes de la réponse aux médicaments.

2. Article 1

Patillon B, Luisi P, Blanché H, Patin E, Cann HM, Génin E, Sabbagh A. *Positive selection in the chromosome 16 VKORC1 genomic region has contributed to the variability of anticoagulant response in humans*. PLoS One. 2012;7(12):e53049. PMID: 23285254

Ce travail a également donné lieu à une brève :
Patillon B, Génin E, Sabbagh A. *La variabilité de réponse aux anticoagulants oraux, une conséquence de la sélection naturelle*. Médecine Sciences, 2013 ;29:159.

OPEN ACCESS Freely available online

PLOS ONE

Positive Selection in the Chromosome 16 *VKORC1* Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans

Blandine Patillon[1,2*☯], Pierre Luisi[3☯], Hélène Blanché[4], Etienne Patin[5], Howard M. Cann[4], Emmanuelle Génin[1¶], Audrey Sabbagh[6¶]

1 Inserm UMRS-946, Genetic Variability and Human Diseases, Institut Universitaire d'Hématologie, Université Paris Diderot, Paris, France, **2** Université Paris Sud, Kremlin-Bicêtre, France, **3** Institute of Evolutionary Biology, CEXS-UPF-PRBB, Catalonia, Barcelona, Spain, **4** Fondation Jean-Dausset-CEPH, Paris, France, **5** Human Evolutionary Genetics, CNRS URA3012, Institut Pasteur, Paris, France, **6** UMR IRD 216, Université Paris Descartes, Paris, France

Abstract

VKORC1 (vitamin K epoxide reductase complex subunit 1, 16p11.2) is the main genetic determinant of human response to oral anticoagulants of antivitamin K type (AVK). This gene was recently suggested to be a putative target of positive selection in East Asian populations. In this study, we genotyped the HGDP-CEPH Panel for six *VKORC1* SNPs and downloaded chromosome 16 genotypes from the HGDP-CEPH database in order to characterize the geographic distribution of footprints of positive selection within and around this locus. A unique *VKORC1* haplotype carrying the promoter mutation associated with AVK sensitivity showed especially high frequencies in all the 17 HGDP-CEPH East Asian population samples. *VKORC1* and 24 neighboring genes were found to lie in a 505 kb region of strong linkage disequilibrium in these populations. Patterns of allele frequency differentiation and haplotype structure suggest that this genomic region has been submitted to a near complete selective sweep in all East Asian populations and only in this geographic area. The most extreme scores of the different selection tests are found within a smaller 45 kb region that contains *VKORC1* and three other genes (*BCKDK, MYST1 (KAT8), and PRSS8)* with different functions. Because of the strong linkage disequilibrium, it is not possible to determine if *VKORC1* or one of the three other genes is the target of this strong positive selection that could explain present-day differences among human populations in AVK dose requirement. Our results show that the extended region surrounding a presumable single target of positive selection should be analyzed for genetic variation in a wide range of genetically diverse populations in order to account for other neighboring and confounding selective events and the hitchhiking effect.

Citation: Patillon B, Luisi P, Blanché H, Patin E, Cann HM, et al. (2012) Positive Selection in the Chromosome 16 *VKORC1* Genomic Region Has Contributed to the Variability of Anticoagulant Response in Humans. PLoS ONE 7(12): e53049. doi:10.1371/journal.pone.0053049

Editor: Yury E. Khudyakov, Centers for Disease Control and Prevention, United States of America

Received September 20, 2012; **Accepted** November 23, 2012; **Published** December 28, 2012

Funding: This work was supported by the Spanish National Institute for Bioinformatics (www.inab.org). PL is supported by a PhD fellowship from "Acción Estratégica de Salud, en el Marco del Plan Nacional de Investigación Científica, Desarrollo e Innovación Tecnológica 2008–2011" from Instituto de Salud Carlos III. BP is supported by a PhD fellowship from the doctoral program in Public Health (ED420: www.ed-sante-publique.u-psud.fr) from Paris Sud University. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing Interests: The authors have declared that no competing interests exist.

* E-mail: blandine.patillon@inserm.fr

☯ These authors contributed equally to this work.

¶ These authors also contributed equally to this work.

Introduction

Oral anticoagulants of antivitamin K type (AVK) − such as warfarin and acenocoumarol − are widely prescribed drugs for the prevention and treatment of arterial and venous thromboembolic disorders [1,2]. They exert their anticoagulant effect by inhibiting the vitamin K 2,3-epoxide reductase complex 1 (VKORC1). Besides well-known physiopathological and environmental factors, including age, sex, body mass index, disease states, co-medications and diet, genetic factors have been identified as major determinants of AVK dose variability [3]. Candidate-gene and genome-wide association studies have identified four main genes − *CYP2C9*, *CYP4F2*, *CYP2C18* and *VKORC1*− which explain together between 28.2% and 43.5% of the AVK dose variance [3,4,5,6,7]. *CYP2C9*, *CYP4F2* and *CYP2C18* encode proteins involved in the hepatic metabolism of AVK [8,9,10]. *VKORC1* encodes the VKORC1 enzyme, which is the direct pharmacologic target of AVK [11,12]. Differences in the worldwide distribution of the most important polymorphisms influencing AVK dosing are likely to underlie the wide interethnic variability in AVK dose requirements: current population-based trends in warfarin dosing, as reported by the International Warfarin Pharmacogenetics Consortium, indicate a mean weekly dose of 21 mg in Asians, 31.5 mg in Europeans and 40 mg in individuals of African ancestry [13].

Recently, Ross *et al.* [14] documented the distribution of four functional variants located in the three main genes known to influence AVK dose requirement − rs9923231 (*VKORC1*), rs1799853 and rs1057910 (*CYP2C9*), and rs2108622 (*CYP4F2*) − in a large set of samples from the Human Genome Diversity Project - Centre d'Etude du Polymorphisme Humain (HGDP-

CEPH) Panel, representing 52 world populations [15]. They observed a pattern of genetic differentiation among human populations for the *VKORC1* single nucleotide polymorphism (SNP) rs9923231. They applied three formal tests of positive selection to the *VKORC1* gene – the locus-specific branch length (LSBL) test [16], the log of the ratio of heterozygosities (ln*RH*) test [17], and Tajima's *D* [18] – using genome-wide data available for the West African, European and East Asian HapMap samples [19]. The tests yielded significant results in the East Asian sample. Interestingly, the rs9923231 SNP (g.-1639G>A), which was found to be a putative target of positive selection [14], is the main genetic determinant of AVK dose requirement and can alone explain between 25% to 30% of the dose variance among patients [4,5,6,7]. This SNP, located in the promoter region, alters a *VKORC1* transcription factor binding site, leading to lower protein expression [20]. By decreasing VKORC1 activity, the derived -1639A allele thus confers an increased AVK sensitivity phenotype and patients carrying one and two -1639A alleles require on average respectively 25% and 50% lower daily warfarin doses than -1639G homozygous carriers to obtain the same anticoagulant effect [21,22]. Understanding the processes of local adaption that may result in high levels of population differentiation and important interethnic differences in the required AVK dose is thus of particular relevance.

During these last few years, newer methods than those proposed by Ross *et al.* have been developed to detect the molecular footprints of positive selection. These methods are particularly well suited to detect classical signatures of selective sweeps, *i.e.* when a new advantageous mutation spreads rapidly to fixation in particular populations (the so-called 'hard sweep' model) [23]. Such a selective sweep occurs too quickly to leave enough time for recombination events to break down the linkage disequilibrium (LD), leading to a similar increase in frequency of alleles at nearby variants. Therefore, the pattern of genetic variation in the genomic region surrounding the selected allele may differ among populations [24], and the selected allele is expected to be carried by a long and frequent haplotype only in those populations that experienced the local adaptive event [25]. Signals of positive selection can thus be detected by looking for an increased genetic differentiation among populations (using methods such as F_{ST} [26] and the Cross-Population Composite Likelihood Ratio (XP-CLR) test [24]), and an extended haplotype homozygosity (EHH) at the putatively selected locus (using methods such as the Cross-Population Extended Haplotype Homozygosity (XP-EHH) test [27] and the integrated Haplotype Score (iHS) [28]). These methods have proved to be powerful and largely complementary to detect and localize a selective sweep, and are more robust to ascertainment bias in SNP discovery than methods based on the allele frequency spectrum such as the Tajima's *D* used by Ross *et al.* [14,29].

In this study, we investigated whether and how positive selection has acted on the *VKORC1* gene locus using these complementary analytic methods. Our first objective was to determine (1) if the selective sweep is restricted to East Asia or if it is detected in other geographic regions, in particular Central South Asia and America, which are geographically close to East Asia, and (2) if it occurred in all East Asian populations or only in a few of them. Thus, we genotyped six *VKORC1* SNPs in the HGDP-CEPH Panel [30] which covers a much wider range of world populations – including 17 populations from East Asia – than the HapMap Panel in which positive selection at the *VKORC1* locus was initially evidenced. Furthermore, by expanding the analysis to a 2 Mb region encompassing the *VKORC1* gene, we sought to determine if the selective sweep identified around *VKORC1* was due to positive selection directly acting on this gene, or if it was caused by positive selection at a nearby linked gene resulting in genetic hitchhiking [23]. Finally, we discuss combining different methods for uncovering distinct selection signatures, in order to both increase power to detect a selective signal and precisely define its genomic location. We address the difficulty, even with such detailed analyses, in identifying the specific target of selection.

Results

VKORC1 Haplotype Study

A haplotype study of the 4.1 kb *VKORC1* gene was carried out with seven *VKORC1* SNPs genotyped in the 52 HGDP-CEPH population samples (Figure 1A). Haplotypes were reconstructed from these SNPs. Seven of these haplotypes had a frequency above 1% in at least one geographic region and were labeled H1 to H7 according to their frequency at the global level (Figure 1B). Four haplotypes are found in at least five geographic regions and only two are shared among all regions. The highest and lowest haplotype diversity values are observed in Sub-Saharan Africa (0.75 ± 0.02) and East Asia (0.19 ± 0.02), respectively. Most individuals carrying the ancestral haplotype (H6), *i.e.* the haplotype carrying the ancestral allele at each SNP, are from Sub-Saharan Africa (Figure 1B and Figure S1). Interestingly, the -1639A allele (rs9923231) conferring the increased sensitivity to AVK is carried by a unique haplotype (H1). This haplotype associated with AVK sensitivity is the most frequent at the worldwide level (49.7%) and shows an extremely high differentiation among geographic regions (Figure 1B). While rare in Sub-Saharan Africa (4.4%), it is found at intermediate frequencies in the Middle East, Europe, Central South Asia, Oceania and America (from 27.8% to 51.2%), and is largely predominant in East Asia (89.6%). The prevalence of H1 tends to be high in all of the 17 East Asian population samples investigated, ranging from 75% in She to 100% in Oroqen (Figure S1). However, the sample size is small for most of them, with 10 or less individuals.

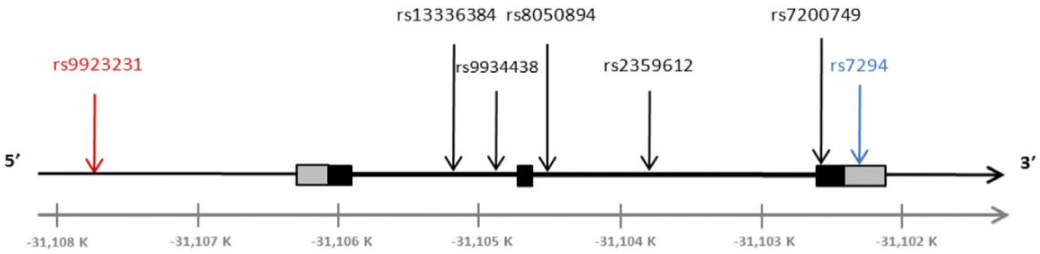
The median-joining haplotype network describes the mutational relationships between the different *VKORC1* haplotypes inferred (Figure 1C). Haplotype H1 differs from the others by two nucleotide substitutions at the functional rs9923231 SNP and at the rs9934438 SNP, which are found in complete LD in all geographic regions ($D' = 1$ and $r^2 = 1$, Figure S2).

Detection of Signatures of Positive Selection

To support the hypothesis that positive selection has played a role in shaping patterns of genetic variation at *VKORC1*, four complementary methods were applied to detect signatures of selective sweeps in the genome. F_{ST} and XP-CLR are both based on allele frequency differentiation, whereas XP-EHH and iHS are based on haplotype structure. Scores for the four test statistics were computed at both the regional and population levels for the seven *VKORC1* SNPs and for some other available SNPs [15] representing the expected neutral genomic background. For each score, a *p*-value was derived from the empirical distribution obtained from the genomic background (*cf.* Material and Methods). We considered as significant any *p*-value below 0.05. The results of the four tests are presented in Table 1 and Table 2.

At the global level, when we evaluated the level of genetic differentiation among the seven HGDP-CEPH Panel geographic regions, an atypical pattern of genetic differentiation was detected for four *VKORC1* SNPs: rs2359612, rs8050894, rs9934438 and rs9923231 ($p<0.05$). The functional rs9923231 polymorphism and the rs9934438 SNP, in complete LD with each other, displayed F_{ST} values falling above the 99th percentile of the empirical

A



B

Haplotype	Haplotypic sequence	Global	AF	ME	EUR	CSA	EA	OCE	AM
		Frequency							
H1	A G A G A G G	0.497	0.044	0.512	0.510	0.309	0.896	0.278	0.413
H2	G G G C G G A	0.335	0.429	0.256	0.307	0.482	0.097	0.722	0.587
H3	G G G C G G G	0.109	0.042	0.177	0.181	0.194	0.004		
H4	G G G C G A A	0.018	0.200	0.016		0.003			
H5	G G G G G G G	0.017	0.108	0.028	0.003	0.005	0.002		
H6	G G G C A G G	0.013	0.128	0.010		0.007			
H7	G A G C G G G	0.004	0.042						

C

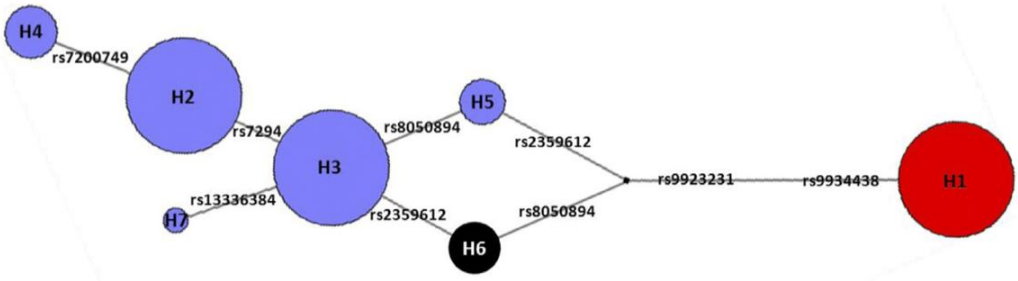


Figure 1. Results of the *VKORC1* haplotype study. (A) Position of the seven SNPs along the *VKORC1* gene. *VKORC1* is a 4.1 kb gene (GenBank accession number AY587020) located at 16p11.2. The three exons of the gene are represented as boxes, with 5′UTR and 3′UTR regions colored in grey and coding regions in black. Flanking and intronic regions are represented as thin and thick lines, respectively. The seven studied SNPs are shown in their sequential order along the *VKORC1* gene. The functional polymorphism rs9923231 located in the promoter, is highlighted in red and the SNP already present in the Illumina 650K chip in blue. Physical position along chromosome 16 is indicated in kb below. **(B) Distribution of *VKORC1* haplotypes at the global and regional level.** For each haplotype, SNPs are listed in the same sequential order than in Figure 1A. Ancestral and derived alleles are shown in blue and orange, respectively. Haplotype labels H1 to H7 were given according to the global haplotype frequency. AF, sub-Saharan Africa; ME, Middle East; EUR, Europe; CSA, Central South Asia; EA, East Asia; OCE, Oceania; AM, America. **(C) Median-joining network of the inferred *VKORC1* haplotypes at the global level.** Circles areas are proportional to the global haplotype frequency and branch lengths to the number of mutations separating haplotypes. Labels of haplotypes are indicated in corresponding circles, and labels of mutations on the network branches. The haplotype carrying the -1639A allele conferring the AVK sensitivity phenotype (H1) is shown in red and the ancestral haplotype (H6) in black.
doi:10.1371/journal.pone.0053049.g001

genome-wide distribution (F_{ST} = 0.32, p = 0.008) (Figure 2A). When global F_{ST} values were computed among the 52 world populations, very similar results were obtained (Table S1). At the inter-regional level, *i.e.* between a given geographic region and the remaining ones, the same four *VKORC1* SNPs showed highly significant F_{ST} values (p<0.01) when comparing Central South Asia and East Asia to the rest of the world (Table 1, Figures 2B and 2C). Regarding East Asia, the highest F_{ST} values (F_{ST} = 0.41, p = 0.003) were also observed for the two SNPs, rs9923231 and rs9934438. For the other geographic regions, no *VKORC1* SNP displayed an inter-regional F_{ST} value as much significant as the ones observed for Central South Asia and East Asia (Table 1 and Figure S3). At the intra-regional level, *i.e.* among populations within a region, no extreme pattern of genetic differentiation (p<0.01) was observed for any *VKORC1* SNP in any geographic region (Table 1 and Figure S4).

The XP-CLR test applied to each geographic region also provided evidence of an atypical pattern of genetic differentiation at the *VKORC1* gene locus, with XP-CLR scores in East Asia ranging from 16.53 (p = 0.050) to 43.44 (p = 0.012) in the 16 kb genomic region centered on *VKORC1* (Table 2). For each of the other six geographic regions, the XP-CLR scores were very low, supporting the existence of a selective sweep restricted to East Asia. In this geographic region, when the XP-CLR test was

Table 1. Results of the inter-regional F_{ST}, intra-regional F_{ST}, XP-EHH and iHS tests in the seven geographic regions.

Region	SNP	DAF[a]	Inter-regional F_{ST}[b]	Inter-regional F_{ST} *p*-value[c]	Intra-regional F_{ST}[d]	Intra- regional F_{ST} *p*-value[c]	XP-EHH score	XP-EHH *p*-value[e]	iHS score	iHS *p*-value[e]
Africa	rs7294	0.63	0.18	0.215	0.13	0.074	−0.94	0.833	−1.24	0.183
	rs7200749	0.20	0.48	0.217	0.02	0.643	−1.50	0.923	−0.31	0.748
	rs2359612	0.82	0.23	0.173	0.09	0.123	−1.15	0.875	−0.96	0.305
	rs8050894	0.16	0.25	0.143	0.09	0.125	−1.14	0.872	0.11	0.909
	rs9934438	0.04	0.36	0.029 *	0.10	0.058	−1.05	0.855	−0.03	0.974
	rs13336384	0.04	0.16	0.329	0.02	0.448	−1.06	0.858	0.14	0.883
	rs9923231	0.04	0.36	0.029 *	0.10	0.058	−1.01	0.845	−0.01	0.989
Middle East	rs7294	0.27	0.02	0.411	0.00	0.906	0.94	0.171	1.55	0.103
	rs7200749	0.02	0.00	0.670	0.02	0.310	1.58	0.069	0.65	0.492
	rs2359612	0.48	0.00	1.000	0.006	0.595	1.17	0.127	2.69	0.009 **
	rs8050894	0.54	0.00	0.849	0.002	0.667	1.15	0.132	−1.40	0.141
	rs9934438	0.51	0.00	0.946	0.01	0.481	1.05	0.149	−1.76	0.066
	rs13336384	0.00	0.005	0.044 *	0.00	1.000	1.07	0.145	NA	NA
	rs9923231	0.51	0.00	0.946	0.01	0.481	1.01	0.156	−1.76	0.066
Europe	rs7294	0.30	0.005	0.670	0.004	0.570	0.94	0.167	0.67	0.474
	rs7200749	0.00	0.02	0.477	0.00	1.000	1.50	0.077	NA	NA
	rs2359612	0.49	0.00	1.000	0.02	0.304	1.15	0.125	2.00	0.039 *
	rs8050894	0.51	0.00	1.000	0.02	0.286	1.14	0.128	−0.98	0.298
	rs9934438	0.51	0.00	0.993	0.02	0.304	1.05	0.145	−1.02	0.281
	rs13336384	0.00	0.005	0.071	0.00	1.000	1.06	0.142	NA	NA
	rs9923231	0.51	0.00	0.993	0.02	0.304	1.01	0.155	−1.02	0.281
Central South Asia	rs7294	0.49	0.06	0.042 *	0.07	0.026*	0.62	0.260	−0.54	0.550
	rs7200749	0.003	0.02	0.261	0.02	0.041*	1.25	0.116	NA	NA
	rs2359612	0.69	0.12	0.002 **	0.07	0.025*	0.89	0.187	1.00	0.280
	rs8050894	0.32	0.12	0.003 **	0.08	0.015*	0.87	0.191	−0.13	0.893
	rs9934438	0.31	0.10	0.006 **	0.07	0.020*	0.78	0.215	−0.20	0.834
	rs13336384	0.00	0.005	0.088	0.00	1.000	0.79	0.210	NA	NA
	rs9923231	0.31	0.10	0.006 **	0.07	0.020*	0.73	0.227	−0.20	0.834
East Asia	rs7294	0.10	0.21	0.063	0.02	0.311	2.68	0.011 *	1.99	0.040 *
	rs7200749	0.00	0.02	0.576	0.00	1.000	3.10	0.005 **	NA	NA
	rs2359612	0.10	0.39	0.005 **	0.02	0.317	2.89	0.008 **	1.92	0.047 *
	rs8050894	0.90	0.38	0.005 **	0.02	0.331	2.88	0.008 **	−1.20	0.200
	rs9934438	0.90	0.41	0.003 **	0.02	0.300	2.81	0.009 **	−1.27	0.174
	rs13336384	0.00	0.005	0.252	0.00	1.000	2.81	0.009 **	NA	NA
	rs9923231	0.90	0.41	0.003 **	0.02	0.300	2.773	0.010 *	−1.274	0.174

Table 1. Cont.

Region	SNP	DAF[a]	Inter-regional F_{ST}[b]	Inter-regional F_{ST} *p*-value[c]	Intra-regional F_{ST}[d]	Intra- regional F_{ST} *p*-value[c]	XP-EHH score	XP-EHH *p*-value[e]	iHS score	iHS *p*-value[e]
Oceania	rs7294	0.72	0.23	0.090	0.00	0.771	0.03	0.456	0.05	0.961
	rs7200749	0.00	0.005	0.401	0.00	1.000	0.51	0.285	NA	NA
	rs2359612	0.72	0.09	0.404	0.00	0.771	0.32	0.346	0.05	0.961
	rs8050894	0.25	0.12	0.327	0.02	0.530	0.29	0.355	0.50	0.584
	rs9934438	0.28	0.08	0.438	0.00	0.749	0.20	0.388	0.50	0.584
	rs13336384	0.00	0.009	0.014 *	0.00	1.000	0.21	0.384	NA	NA
	rs9923231	0.28	0.08	0.438	0.00	0.749	0.16	0.404	0.50	0.584
America	rs7294	0.58	0.11	0.320	0.17	0.195	0.76	0.207	NA	NA
	rs7200749	0.00	0.01	0.553	0.00	1.000	1.15	0.125	NA	NA
	rs2359612	0.59	0.02	0.674	0.17	0.190	0.96	0.162	NA	NA
	rs8050894	0.41	0.02	0.667	0.17	0.190	0.95	0.163	NA	NA
	rs9934438	0.41	0.01	0.743	0.17	0.190	0.88	0.178	NA	NA
	rs13336384	0.00	0.006	0.013 *	0.00	1.000	0.89	0.176	NA	NA
	rs9923231	0.41	0.01	0.743	0.17	0.190	0.85	0.185	NA	NA

[a]Derived allele frequency estimated at the global level.
[b]F_{ST} estimated at the inter-regional level, *i.e.* between a given geographic region and the remaining ones.
[c]*p*-values are derived from the genome-wide empirical distribution of F_{ST} values.
[d]F_{ST} estimated at the intra-regional level, *i.e.* among populations within a region.
[e]*p*-values are derived from the empirical distribution of the iHS and XP-EHH scores along the chromosome 16.
*$p<0.05$; ** $p<0.01$; *** $p<0.005$.
NA: Not Applicable (for iHS: when a gap>200 kb between successive SNPs is found in the region in the region delimited by the SNPs where the EHH value drops below 0.05 around the core SNP).
doi:10.1371/journal.pone.0053049.t001

Table 2. Results of the XP-CLR test in a 16 kb region centered on *VKORC1* in the seven geographic regions.

Region	Physical position	XP-CLR score	XP-CLR *p*-value[a]
Africa	31005354	0.00	1.000
	31009354	0.00	1.000
	31013354	0.96	0.289
	31017354	0.58	0.348
	31021354	0.09	0.470
Middle East	31005354	4.00	0.138
	31009354	0.85	0.306
	31013354	3.28	0.158
	31017354	0.27	0.403
	31021354	6.25	0.092
Europe	31005354	0.54	0.351
	31009354	0.00	1.000
	31013354	2.63	0.186
	31017354	0.15	0.427
	31021354	2.40	0.198
Central South Asia	31005354	0.03	0.464
	31009354	0.00	1.000
	31013354	0.00	1.000
	31017354	0.01	0.476
	31021354	0.00	0.490
East Asia	31005354	24.08	0.032 *
	31009354	16.53	0.050 *
	31013354	30.49	0.023 *
	31017354	26.82	0.028 *
	31021354	43.44	0.012 *
Oceania	31005263	0.00	1.000
	31009263	0.00	1.000
	31013263	0.00	1.000
	31017263	0.00	1.000
	31021263	0.00	1.000
America	31005354	0.00	1.000
	31009354	0.00	1.000
	31013354	0.00	1.000
	31017354	0.01	0.587
	31021354	0.00	0.597

[a]*P*-values are derived from the empirical distribution of the XP-CLR scores along the chromosome 16.
*$p<0.05$; ** $p<0.01$; *** $p<0.005$.
doi:10.1371/journal.pone.0053049.t002

performed for each population, all of the 17 population samples, except Oroqen, showed this extreme pattern of genetic differentiation, with at least three significant XP-CLR scores out of the five scores computed in the 16 kb genomic region surrounding *VKORC1* (Table S2). As most of the SNPs in the *VKORC1* genomic region have reached fixation in the Oroqen sample, XP-CLR scores could be calculated for only very few SNPs on either side of *VKORC1*, making difficult the interpretation of XP-CLR results in this sample.

Regional results obtained with the extended haplotype-based XP-EHH test indicated that the unusual pattern of genetic differentiation observed at the *VKORC1* gene locus resulted from a selective sweep in East Asia. Significant XP-EHH scores, ranging from 2.68 ($p=0.011$) to 3.10 ($p=0.005$), were observed for the seven *VKORC1* SNPs in East Asia, while no significant values were observed for any other geographic region (Table 1). For East Asian populations, evidence for a selective sweep was detected in all 17 population samples with significant XP-EHH scores for each of the seven *VKORC1* SNPs, ranging from 1.84 ($p=0.049$) in the Dai sample for rs7294, to 3.78 ($p=0.004$) in the Tujia sample for rs8050894 (Table S3).

With the iHS test, only two *VKORC1* SNPs (rs7294 and rs2359612) exhibited significant iHS scores in East Asia ($p=0.040$ and 0.047, respectively; Table 1). Two other significant scores were observed for the rs2359612 SNP in the Middle East (2.69, $p=0.009$) and Europe (2.00, $p=0.039$). At the population level in East Asia, only three samples (Hezhen, Lahu, and Yakut) displayed significant iHS scores for two, three and four SNPs, respectively (Table S3).

The four selection tests consistently evidenced the signature of a selective sweep involving the *VKORC1* genomic region in East Asia. However, this result did not allow us to determine with certainty that *VKORC1* is the direct target of positive selection. A linked gene could be the target instead, resulting in genetic hitchhiking of *VKORC1* [23]. In an attempt to seek the true target of positive selection, we probed the downloaded chromosome 16 genotypes [15] with the four tests for selection and examined the results over an extended 2 Mb genomic region centered on *VKORC1*. We focused on clusters of selection test scores with highly significant *p*-values ($p<0.01$) for East Asia only. Three clusters were observed (Figure 3): (i) ~ 570 kb downstream of *VKORC1*, the first cluster was found with partially overlapping clusters of extreme XP-CLR and XP-EHH scores over a region of 64 and 39 kb, respectively, involving the genes *ITGAL*, *ZNF768*, and *ZNF747*; (ii) at or close to *VKORC1* genomic position, the second cluster was determined by overlapping clusters of extreme F_{ST} values when comparing East Asia to the rest of the world (with the lowest *p*-values observed for the same two *VKORC1* SNPs evidenced before, rs9923231 and rs9934438) and extreme XP-CLR and XP-EHH scores. These clusters ranged in size from 45 to 244 kb; (iii) ~ 230 kb upstream of *VKORC1*, the third cluster of 32 kb was found with XP-EHH and concerned the genes *ITGAM* and *ITGAX*. If SNPs within clusters are in high LD ($D'\geq0.97$, except for one SNP in the third cluster), only limited LD exists between the SNPs located in the different clusters (Figure 4 and Figure S5) and several recombination hotspots are present between these clusters (Figure 4). This suggests that each of the three clusters represents a different adaptive event.

Examination of the second cluster showed that *VKORC1* is contained in a block of strong LD spanning ~ 505 kb in East Asia (Figure 4 and Figure S5). Similar LD blocks were observed for Central South Asia and Europe, and to a lesser extent, for the Middle East (Figure S5). This LD block encompasses 25 genes (Figure 4). We used the most extreme F_{ST}, XP-CLR and XP-EHH scores in order to spatially localize a target of selection within the LD block. Significant XP-CLR scores ($p<0.05$) were found in a 350 kb region encompassing 19 genes including *VKORC1* (Table S4). XP-EHH scores were almost all significant at the 0.05 threshold but four adjacent genes *VKORC1*, *BCKDK*, *MYST1* (*KAT8*) and *PRSS8* displayed most extreme XP-EHH scores ($p<0.01$). Clusters of highly significant F_{ST} values when comparing East Asia to the rest of the world ($p<0.01$) and significant global F_{ST} values ($p<0.05$) were also found for these four genes (Table S5). It is thus probable that the selective pressure has targeted one of these genes.

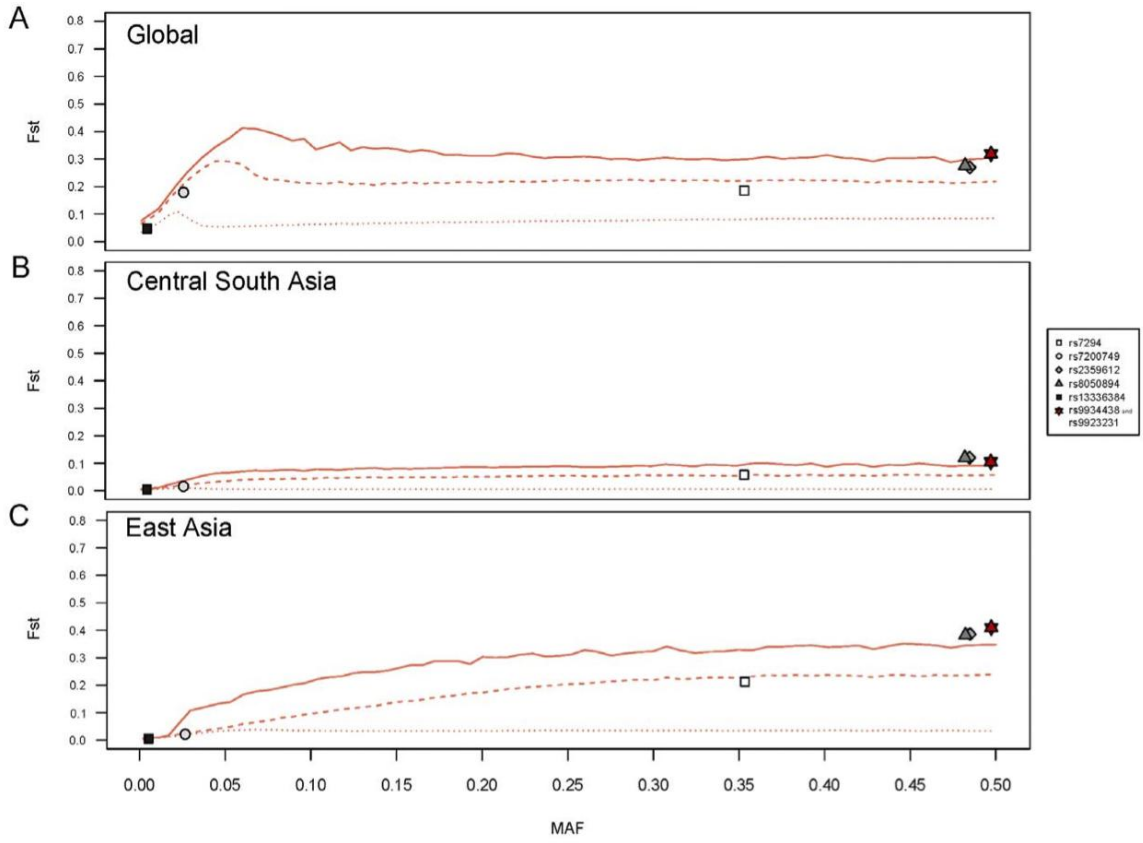


Figure 2. Atypical patterns of genetic differentiation observed for *VKORC1* SNPs. Genome-wide empirical distributions of F_{ST} values were constructed from 644,143 SNPs having a MAF ≥0.001 at the global level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against their global MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.
doi:10.1371/journal.pone.0053049.g002

When did the -1639A VKORC1 Allele begin to Increase in East Asia?

The time at which the frequency of the -1639A allele started to increase in East Asia was estimated by using a maximum-likelihood method [31] with the 17 East Asian HGDP-CEPH sample data. Our analysis yielded an age estimate of 181 generations (95% CI: 128–256 generations). Assuming a generation time of 25 years, the expansion therefore occurred about 4,525 years ago (95% CI: 3,200–6,400 years).

Discussion

Numerous genes involved in absorption, distribution, metabolism and excretion (ADME) of drugs, exhibit evidence of recent positive selection and/or high population differentiation levels [32]. However, there are fewer examples of the action of natural selection on genes involved in the pharmacodynamics of drugs, such as *VKORC1*. Although numerous surveys have examined the genetic polymorphism of *VKORC1* in samples from diverse ethnic origins [13,20,33,34,35,36,37], these studies provided an incomplete picture of haplotype diversity because different sets of SNPs were used and worldwide coverage was incomplete. In this study, we took advantage of the worldwide coverage of the HGDP-CEPH Panel to provide the first detailed analysis of *VKORC1* population diversity using the same set of SNPs. Haplotype analysis revealed that the -1639A derived allele that confers AVK sensitivity is carried by a unique haplotype in all 52 population samples investigated. This haplotype associated with AVK sensitivity is predominant in East Asia, rare in Sub-Saharan Africa and occurs at intermediate frequencies in other geographic regions. Because it is found in Sub-Saharan Africa and other world populations, this haplotype is probably rather old. Its geographic distribution leads to striking differences between East Asian and non East Asian samples for genetic susceptibility to AVK sensitivity.

One explanation for worldwide diversity of this haplotype could be positive selection. This hypothesis was supported by five genome-wide scans that found atypical patterns of the allele frequency spectrum [38], extended LD [39,40], and unusual genetic differentiation [40,41,42] in a 450 kb genomic region encompassing *VKORC1*. When specified, the target population was Asian [38,40]. Ross *et al.* [14] found evidence of positive selection at *VKORC1* in the East Asian HapMap sample, based on the level of genetic diversity (ln*RH* test [17]), genetic differentiation (LSBL test [16]) and allele frequency spectrum (Tajima's *D* [18]).

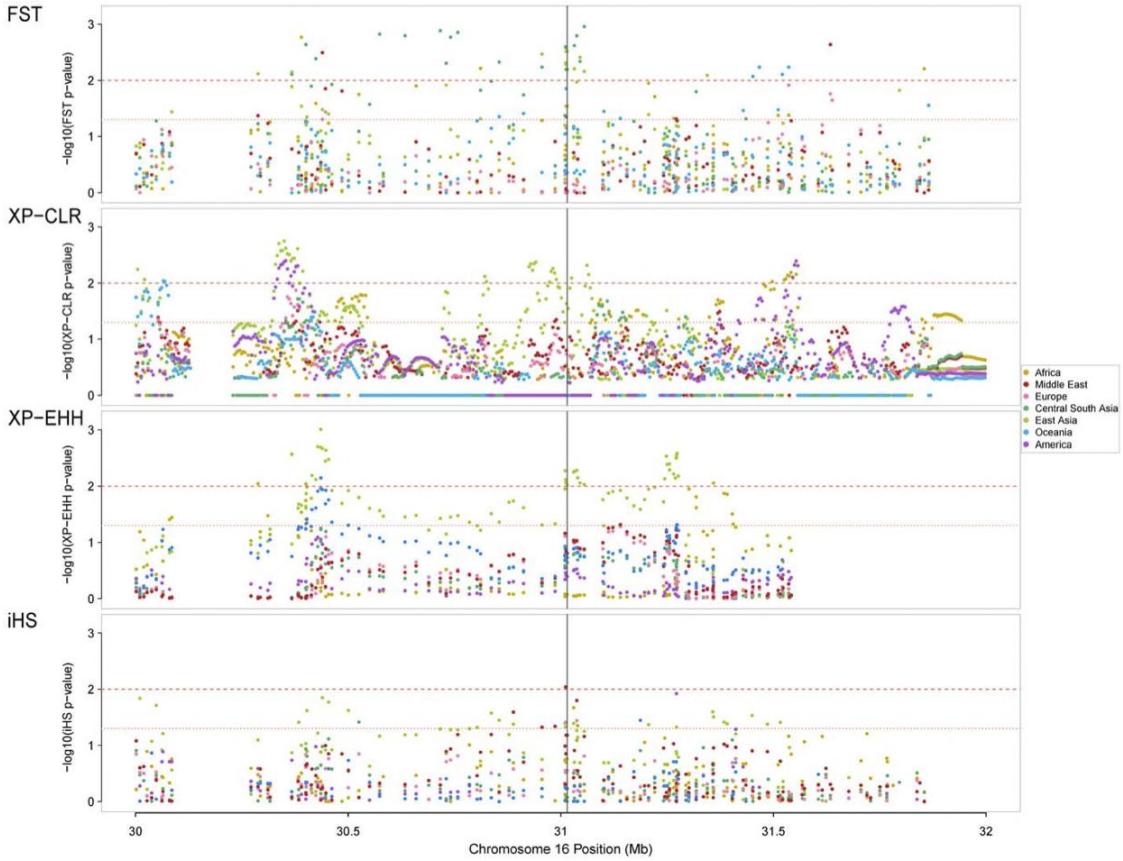


Figure 3. Distribution of $-\log_{10}$ (*p*-values) for four selection tests across a 2 Mb region centered on *VKORC1*. A black vertical line indicates the physical position of *VKORC1* on chromosome 16. Horizontal red dotted and dashed lines show 0.05 and 0.01 chromosome-wide significance levels, respectively. The selection tests (inter-regional F_{ST}, XP-CLR, XP-EHH and iHS, respectively) were separately applied in each of the seven geographic regions.
doi:10.1371/journal.pone.0053049.g003

In this study, we provided compelling evidence of positive selection at the *VKORC1* gene locus in East Asia and only in this geographic region. A footprint of natural selection was found in each of the widely distributed 17 HGDP-CEPH East Asian population samples. By using four different tests of positive selection and by assessing significance at a given locus on the basis of an empirical distribution derived from the genomic background, we believe we can be confident that positive selection, rather than demographic forces, accounts for the data presented here. Indeed, it is well known that large allele frequency differences between populations are not infallible proofs of positive selection: these can also result from genetic drift, migration and other neutral demographic processes [43,44]. This might be the explanation for the significant inter-regional F_{ST} values observed in Central South Asia (Table 1 and Figure 2B).

Because the XP-EHH test is designed to detect fixation events that are relatively young (~ 30,000 years) [27], the selective event we have detected is likely to be rather recent. This is indeed supported by an age estimate of 4,525 years (95% CI: 3,200–6,400 years) for the time at which the *VKORC1* -1639A allele started to increase in frequency in East Asia. The poor performance of the iHS test that detected only very few signals of positive selection in this study could have been predicted since its power to detect selective sweeps involving alleles near fixation is known to be low [28,45]. By contrast, XP-EHH and XP-CLR perform better when the allele targeted by selection is near fixation and indeed showed strong evidence of a selective sweep in this study [24,27].

In an attempt to determine if the *VKORC1* gene has been the direct target of positive selection or if it reflects genetic hitchhiking [23], we extended our analysis to a 2 Mb region surrounding the *VKORC1* gene (Figure 3). Apart from the highly significant footprint of positive selection localized in the *VKORC1* region, two other significant signals, at ~ 570 kb downstream and ~ 230 kb upstream of *VKORC1*, were detected with XP-CLR and/or XP-EHH in East Asia. These two regions contains genes that belong to the same integrin family – specifically to the CD11 gene cluster: *ITGAL* downstream, and adjoining genes *ITGAM* and *ITGAX* upstream – involved in immune functions and being thus good candidates for positive selection [46,47,48]. However, since SNPs located in these integrin genes show limited LD with those of *VKORC1*, a single adaptive event is unlikely. Apart from East Asia, the *ITGAL* region showed signals of positive selection in other

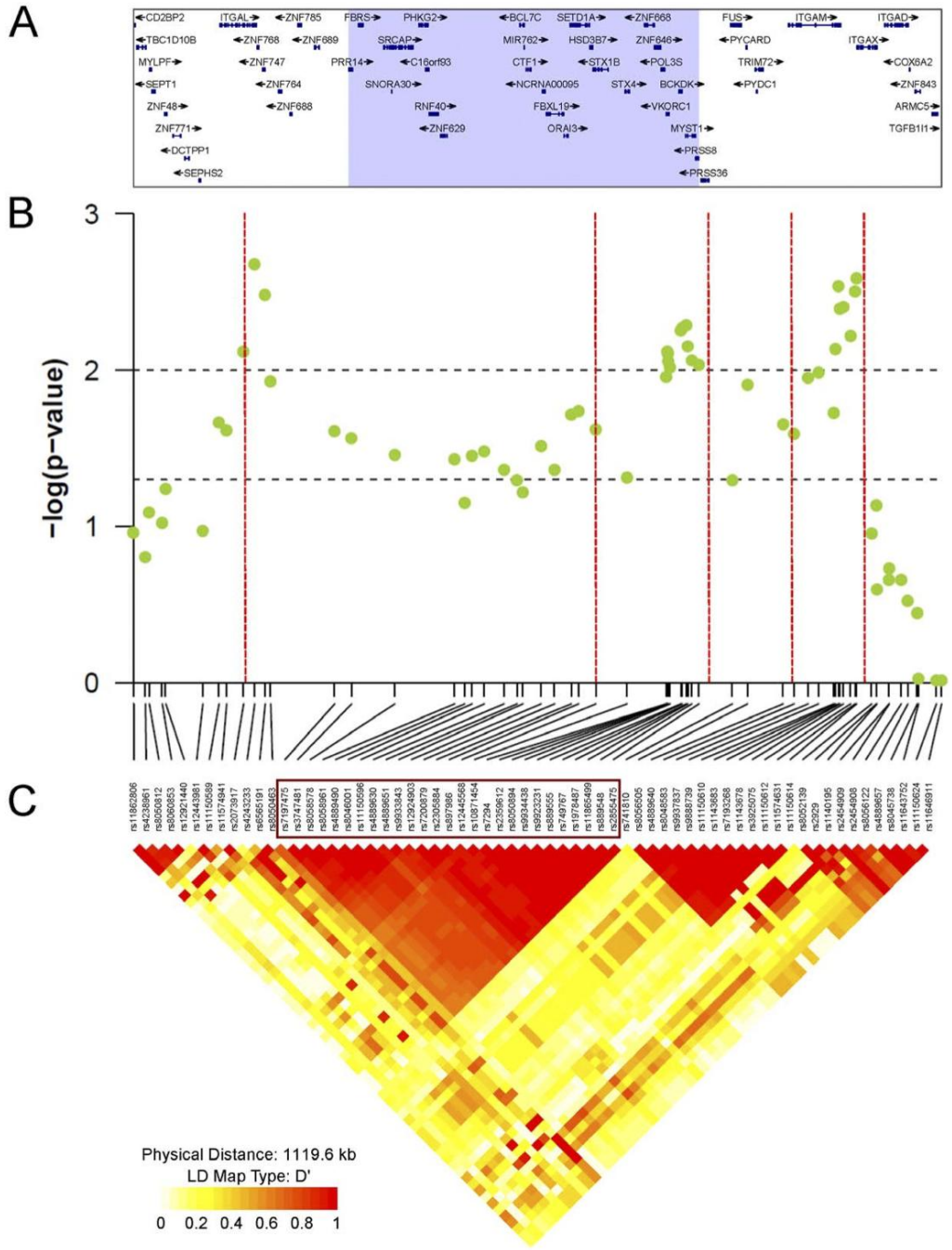


Figure 4. Detailed analysis of a 1.1 Mb genomic region surrounding the *VKORC1* gene locus in East Asia. The boundaries of the region displayed (chr16:30,271,572-31,391,123; UCSC human genome build hg18) were chosen so as to include the three clusters of significant scores detected in East Asia by the selection tests in the 2 Mb region centered on *VKORC1* (Figure 3). (**A**) **Name and location of genes.** Exons are displayed as blue boxes and the transcribed strand is indicated with an arrow. Genes located in the block of strong LD encompassing *VKORC1* and including the SNPs in the red box shown in Figure 4C, are highlighted in the grey area. (**B**) **XP-EHH results in East Asia.** The significance of the XP-EHH scores ($-\log_{10}$ empirical *p*-value) are shown for individual SNPs with a MAF ≥0.01 in East Asia. Horizontal dashed lines indicate 0.05 and 0.01 chromosome-wide significance levels. Recombination hotspots detected in HapMap Phase II data are indicated by red vertical dotted lines. The data

and methods used to derive these hotspots are available from the HapMap website (http://www.hapmap.org/) [83,84]. (**C**) **LD plot.** Pairwise LD values, depicted as *D'*, are shown for SNPs with a MAF ≥0.01 in East Asia. *D'* values are displayed in different colors from yellow to red for *D'* = 0 to *D'* = 1, respectively. The red box highlights SNPs included in the LD block encompassing *VKORC1*. The plot was produced using the snp.plotter R package [74].
doi:10.1371/journal.pone.0053049.g004

geographic regions (America with XP-CLR, and Sub-Saharan Africa and Oceania with XP-EHH), arguing for a different evolutionary history from that of *VKORC1*, which was only found in East Asia. This observation emphasizes the need for studying the geographic distribution of a selective event in a wide range of genetically diverse populations, as per Scheinfeldt *et al.* [49] who, after performing a detailed analysis of a 3 Mb region surrounding a gene showing strong footprints of positive selection, discovered patterns of genetic variation consistent with the presence of a cluster of three independent selective events occurring in different populations. By extending their analysis to the entire genome, they identified several other genomic regions exhibiting evidence for the presence of multiple and independent selective targets, suggesting that clusters of adaptive evolution, such as the one detected herein, are widespread in the human genome.

After delimitating the selective signal for *VKORC1* by analyzing selective events identified in the 2 Mb region just described, we aimed at precisely mapping the gene targeted by positive selection. *VKORC1* is located in a ~ 505 kb LD block in East Asia containing 25 genes (Figure 4), and the selective pressure could have targeted any gene in this LD block. We used F_{ST}, XP-CLR and XP-EHH scores to spatially localize possible targets of positive selection within the LD region. A block of four adjacent genes – *VKORC1, BCKDK, MYST1,* and *PRSS8*– was found to be the most likely selective target (Table S4).

BCKDK codes for the mitochondrial branched chain ketoacid dehydrogenase kinase. *MYST1* and *PRSS8* are two immunity-related genes, listed as candidates for positive selection in several databases [40,42,50]. If, indeed, one of these three genes is the target of the selective sweep detected here, it should contain a functional variant of high frequency in East Asia and we did not find such a variant in HapMap data.

Assuming that selection has directly targeted the *VKORC1* gene, the advantage would then probably be related to vitamin K metabolism,vitamin K being the only known substrate of VKORC1. This vitamin plays a crucial role in the synthesis of vitamin K-dependent (VKD) proteins, especially blood coagulation factors, which requires VKORC1 activity [51,52]. Large geographic differences in dietary vitamin K intake, especially in vitamin K2, exist between human populations, with the highest plasma levels found in Asian populations, as compared to Europeans and Africans [53,54]. These differences could be explained by the wide consumption of fermented soybean food (*natto*) - a major source of vitamin K2 - in East Asia [55,56]. It is then possible that, at some points in the history of East Asian populations, these high levels of vitamin K intakecould have been deleterious and created a selective pressure against *VKORC1* gene expression and coagulant activity. There is, however, no report so far of a deleterious effect associated with a high consumption of vitamin K and it is more the low dietary vitamin K intake that is problematic, hampering the adequate synthesis of VKD proteins in extrahepatic tissues notably bone and arterial vessels [57]. An alternative hypothesis could be that a naturally occurring environmental molecule of AVK type - such as a coumarin derivative - specifically found in East Asia, exerted a selective pressure on the *VKORC1* gene in populations of this region during their recent history. Such molecules are present in the nature, as illustrated by the example of the sweet clover disease that affected cattle in Canada and North America in the 1920's. Sweet clover hay, used to feed cattle, contains a natural coumarin that is oxidized in mouldy hay to form dicoumarol, a hemorrhagic agent. Its discovery led to the synthesis of coumarin derivatives used in clinical application as oral anticoagulants since the 1940's [58,59]. Evidence of an effect of warfarin in shaping *VKORC1* genetic diversity could be found in rats and mices. Indeed, since the introduction in the 1950's of this molecule as rodenticide, mutations in the *VKORC1* gene conferring warfarin resistance have spread in rodent populations but the mechanisms by which they lead to warfarin resistance are still not elucidated [60,61,62,63].

In conclusion, we found that the *VKORC1* genomic region exhibits diversity patterns consistent with the action of positive selection in East Asia. Nearly complete selective sweeps, such as the one described herein, are believed to be rare in recent human adaptive history [64,65,66,67]. This selective event is probably responsible for the spread of the derived -1639A allele conferring the increased AVK-sensitive phenotype in East Asian populations and contributes to present-day differences among human populations in the genetic sensitivity to AVK. A detailed analysis of the extended *VKORC1* genomic region revealed selective signals at several independent genetic loci, indicating a complex evolutionary history for this chromosome 16 region. Our evolutionary analysis emphasizes the importance of considering the surrounding genomic region of a candidate gene for selection in order to avoid erroneous conclusions about the true target of selection. We show here that the gene targeted by selection could be either *VKORC1* or another gene located in the 45 kb region covered by selective sweep detected in East Asia. Our ability to identify the target of selection may be limited by the number of genetic polymorphisms investigated. Examining the selective signal with more genetic variation using whole-genome sequences from the 1000 Genomes Project [68] may well improve the mapping of the gene targeted by selection. Furthermore, allele frequency spectrum bias tends to be minimized with whole genome sequences, which may allow the use of tests for natural selection based on this spectrum.

Materials and Methods

The HGDP-CEPH Panel

We used the HGDP-CEPH Panel that presently includes 1,064 individuals from 52 populations worldwide [69]. For the analysis presented here, the standardized subset panel H952 containing no first nor second degree relative pairs, was used [70]. This subpanel includes 952 individuals grouped into seven broad geographic regions as defined by Li *et al.* [15]: Sub-Saharan Africa (N = 105), the Middle East and Mozabites from north Africa (N = 163), Europe (N = 158), Central South Asia (N = 202), East Asia (N = 232), Oceania (N = 28) and America (N = 64). A full description of the 52 samples included in the HGDP-CEPH Panel is provided in Table S6.

SNP Genotyping

A total of 940 individuals from the original H952 subpanel were previously genotyped by Li *et al.* [15] with the Illumina HumanHap 650 K platform and their genotypes at 644,258 autosomal SNPs were downloaded from the public HGDP-CEPH

database (http://www.cephb.fr/en/hgdp/). Only one SNP (rs7294) from this dataset is located in the *VKORC1* gene. We additionally genotyped six SNPs in *VKORC1* in the 940 individuals, using the TaqMan® SNP Genotyping Assay-by-Design method in 5 µl reaction volumes according to the manufacturer's protocol (Applied Biosystems, Foster City, CA): rs9923231 (g.-1639G>A) located in the promoter region, rs13336384 and rs9934438 in the first intron, rs2359612 and rs8050894 in the second intron, and rs7200749 in the third exon (Figure 1A). Missing genotype rates varied from 0.5% to 2.2% for SNPs rs7200749 and rs9934438, respectively. Since the two SNPs rs9923231 and rs9934438 were found in complete LD in the seven geographic regions (Figure S2), we were able to impute the missing genotypes of a given SNP using available information from the other, leading to a total of 0.96% missing genotypes for these two SNPs. No significant deviations from the Hardy-Weinberg proportions were observed for any *VKORC1* SNP in any of the 52 population samples at the 0.01 significance level (data not shown). Allele frequency distributions of the seven *VKORC1* SNPs in the 52 population samples are shown in Figure S6.

Statistical Analysis

***VKORC1* haplotype study.** To investigate the worldwide diversity of the *VKORC1* gene, we conducted a haplotype study using the seven genotyped SNPs. A total of 931 individuals with less than three missing genotypes were included in the haplotype reconstruction. For each geographic region, haplotype frequencies were estimated with the Bayesian statistical method implemented in Phase v2.1 [71] using defaults parameters. To avoid the convergence of the algorithm to a local maximum, we ran it 10 times with different random seeds and kept the output from the run with the best average value. The worldwide haplotype frequencies were then calculated as the weighted average of the frequencies estimated in each of the seven geographic regions. Similar results were obtained when a single pooled sample of all individuals was considered in the haplotype frequency estimation (data not shown). Since information on ancestral allele state is required to distinguish between ancestral and derived haplotypes, we used the snp131OrthoPt2Pa2Rm2.txt file downloaded from the UCSC genome browser (http://genome.ucsc.edu/) which provides the orthologous alleles in chimpanzee, orangutan and rhesus macaque. For each SNP, the allele shared by the three species was identified as the ancestral allele. Haplotype networks were drawn with the Network v4.5.1.6 software (http://www.fluxus-engineering.com/), using the median-joining algorithm which builds the minimum spanning network from the given haplotypes by favoring short connections [72]. LD analyses were performed with Haploview v4.1 [73] and the snp.plotter R package [74], using Lewontin's disequilibrium coefficient D' [75] and the correlation coefficient r^2 [76].

Detection of Signatures of Positive Selection

To explore whether *VKORC1* has evolved under positive selection in humans, we looked for two distinct genetic patterns of a selective sweep that are expected to remain detectable in the genome over different time scales after the action of natural selection: (i) an important genetic differentiation among populations nearby the locus of interest, and (ii) the presence of unusually frequent and long haplotypes in the surrounding genomic region. For each method, we used an outlier approach to calculate the p-values of the computed scores. Under this approach, an empirical distribution is constructed using other SNPs in the genome that are assumed to be neutral and to represent the genomic background under neutrality. An empirical p-value is computed that corresponds to the proportion of values from the empirical distribution that are higher than the value observed at the locus of interest. If the value obtained for the SNP of interest is greater than the 95th percentile ($p<0.05$) of the empirical distribution, positive selection is invoked. For that purpose, we used the empirical distributions obtained from the scores calculated either on a genome-wide (all autosomal chromosomes) or chromosome-wide (chromosome 16, where *VKORC1* is located) basis.

First, we used two statistics, F_{ST} and XP-CLR, which measure the genetic differentiation among human populations [24,26]. These methods are able to detect selective sweeps that have occurred up to 75,000 years ago [77]. The fixation index F_{ST} [78] quantifies the proportion of genetic variance explained by allele frequency differences among populations. F_{ST} ranges from 0 (for genetically identical populations) to 1 (for completely differentiated populations). We calculated F_{ST} values using the BioPerl module PopGen [79] for each autosomal SNP with a minor allele frequency (MAF) $\geq 10^{-3}$ (644,143 SNPs) at three different levels: (i) global level (either among the seven HGDP-CEPH Panel geographic regions or among the 52 Panel populations), (ii) inter-regional level (each geographic region versus the remaining ones), and (iii) intra-regional level (among populations within a region). Since F_{ST} strongly correlates with heterozygosity [41,80,81], empirical p-values were calculated within bins of 10,000 SNPs grouped according to MAF. The resulting distributions represent the average genetic differentiation of human populations corrected for heterozygosity.

We next applied the XP-CLR test [24] which identifies selective sweeps in a population by detecting significant genetic differentiation in an extended genomic region of interest as compared to a reference population. This method presents both the advantages of being robust to ascertainment bias and of not requiring any information on haplotypes, thus avoiding errors of haplotype estimation from genotype data. XP-CLR scores were computed at regularly spaced grid points (every 4 kb) across chromosome 16 using the genotypes from SNPs within overlapping windows of 0.1 cM around each grid point. To account for different SNP densities among genomic regions, we restricted to 200 the maximal number of SNPs used to compute a XP-CLR score within the 0.1 cM genomic region, by removing excess SNPs at random. We applied this method by considering all SNPs with a MAF $\geq 10^{-3}$ on chromosome 16 at both the regional and population levels (17,729 SNPs). P-values were calculated from the empirical distribution of the collected scores obtained with these SNPs. XP-CLR requires the definition of a reference population: the Sub-Saharan African samples were used as a reference for non Sub-Saharan African regions, and the European samples as a reference for Sub-Saharan Africa. For the analyses performed at the population level, we defined the Yoruba as the reference for non Sub-Saharan African samples, and the French for Sub-Saharan African samples.

The second class of methods that we used is based on EHH, *i.e.* the sharing of identical alleles across relatively long distances by most haplotypes in population samples [25]. In brief, the EHH is computed for a given SNP (the core SNP) of a sequence being interrogated for a selective sweep. In the absence of a selective sweep, recombination events break down haplotypes relatively rapidly with time and with increasing distance from the core SNP. In the case of a selective sweep, LD tends to maintain the haplotype carrying the selected allele, and the relative frequency of this (favored) haplotype will increase with time leading to so-called EHH. Integration of genetic distance in both directions from the core SNP can be used to discriminate between selected and non-selected alleles, and be applied to ancestral and derived alleles.

Analytic methods based on EHH are able to detect recent selective sweeps (*i.e.* those occurring less than 30,000 years ago [77]). Such analyses require haplotype data. We used fastPHASE v1.3.0 EM algorithm [82] to infer haplotypes with chromosome 16 SNPs for individuals from each geographic region. For each region, the *K*-selection procedure was first run several times in order to define the optimal number of clusters of similar haplotypes by minimizing chance error rates. Ultimately, phase was determined with $K=6$ for Oceania, $K=14$ for Europe and Central-South Asia and $K=12$ for the remaining regions. Using these values, the EM algorithm was then run with 20 random starts and 25 iterations.

Once haplotypes were reconstructed, we computed the XP-EHH statistic [27] that compares the integrated EHH computed in a test population versus that of a reference population. Therefore, this method detects a sweep in which the selected allele has risen to near fixation in one population but remains polymorphic in the other. XP-EHH scores were computed using the same parameters as those described in Sabeti *et al.* (2007). Reference populations were defined as for XP-CLR.

We finally applied the iHS [28] that compares the rate of EHH decay observed for both the derived and ancestral allele at the core SNP. An extremely positive or negative value at the core SNP provides evidence of positive selection with unusually long haplotypes carrying the ancestral or the derived allele, respectively. The raw iHS scores were computed using the iHS option implemented in the WHAMM software developed by Voight *et al.* (2006). The scores were standardized to have null mean and unit variance in 5% bins of the derived allele frequency at the core SNP. Information on ancestral allele state was obtained from the snp131OrthoPt2Pa2Rm2.txt file downloaded from the UCSC website. We were unable to determine with certainty the ancestral allele status of 111 SNPs on chromosome 16 and we removed them from the analysis.

XP-EHH and iHS scores were calculated for all available SNPs on chromosome 16 (19,733 and 19,622, respectively) at both the regional and population levels. The resulting distributions were used to calculate empirical *p*-values.

The genetic map used for applying XP-CLR, XP-EHH and iHS was retrieved from release 22, build 36 of HapMap (www.hapmap.org).

Age of the Expansion of the -1639A *VKORC1* Allele in East Asia

We inferred the age at which the -1639A allele started to increase in frequency in East Asia by estimating the age of the most recent common ancestor carrying this allele in East Asia using the likelihood-based method implemented in the Estiage program [31]. This method assumes that all individuals derive from a common ancestor who introduced the mutation *n* generations ago. Estimation of *n* is based on the length of the haplotype shared by the individuals, which is estimated through the identification of recombination events on the ancestral haplotype by taking into account allele frequencies and recombination rates. We estimated *n* using only one haplotype per East Asian population sample (*i.e.*, 17 haplotypes). For each population, this one haplotype was constructed by taking at each locus over a 6 Mb region the allele the most frequently seen in individuals from the population carrying the -1639A allele. A mutation rate of 10^{-6} per individual and per generation, and a 25-year generation time were assumed.

Supporting Information

Figure S1 Distribution of *VKORC1* haplotypes in the 52 HGDP-CEPH samples. The haplotype carrying the -1639A allele (H1) is represented in red and the ancestral haplotype (H6) in black.
(TIF)

Figure S2 Pairwise LD between the seven *VKORC1* SNPs at the regional and global level. Red squares indicate statistically significant (logarithm of odds >2) LD between the pair of SNPs, as measured by the *D'* statistic [75] with the Haploview software [73]; darker colors of red indicate higher values of *D'*, up to a maximum of 1. White squares indicate pairwise *D'* values of <1 with no statistically significant evidence of LD. Blue squares indicate pairwise *D'* values of 1 but without statistical significance.
(TIF)

Figure S3 Genome-wide empirical distributions of inter-regional F_{ST} values against MAF in the seven geographic regions. Empirical distributions of F_{ST} were constructed by calculating an F_{ST} value for 644,413 SNPs having a MAF ≥0.001 at the global level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against their global MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.
(TIFF)

Figure S4 Genome-wide empirical distributions of intra-regional F_{ST} values against MAF in the seven geographic regions. Empirical distributions of F_{ST} were constructed by calculating an F_{ST} value for all SNPs having a MAF ≥0.001 at the intra-regional level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against the regional MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.
(TIFF)

Figure S5 LD patterns over a 2 Mb region centered on *VKORC1* in the seven geographic regions. Pairwise LD, depicted as *D'*, is shown for SNPs with a MAF ≥0.05 at the global level. *D'* values are displayed in different colors from yellow to red for $D'=0$ to $D'=1$, respectively. The plot was produced using the snp.plotter R package [74]. The vertical dashed lines delineate *VKORC1* gene position.
(TIF)

Figure S6 Allele frequency distribution of the seven *VKORC1* SNPs in the 52 HGDP-CEPH samples: rs9923231 **(A)**, rs13336384, **(B)** rs9934438 **(C)**, rs8050894 **(D)**, rs2359612 **(E)**, rs7200749 **(F)** and rs7294 **(G)**. The derived and ancestral alleles are represented in orange and blue, respectively.
(TIF)

Table S1 Global F_{ST} values among populations and among regions for the seven *VKORC1* SNPs.
(XLS)

Table S2 Results of the XP-CLR test in a 16 kb region centered on *VKORC1* in the 52 HGDP-CEPH samples.
(XLS)

Table S3 Results of the XP-EHH and iHS tests in the 52 HGDP-CEPH samples.
(XLS)

Table S4 Results of the XP-CLR test in the ~ 500 kb genomic region of the LD block encompassing *VKORC1* in East Asia.
(XLS)

Table S5 Results of the XP-EHH, iHS tests, inter-regional F_{ST} and global F_{ST} for all SNPs located in the linkage disequilibrium block encompassing *VKORC1* in East Asia.
(XLS)

Table S6 Description of the 52 HGDP-CEPH samples grouped into seven main geographic regions.
(XLS)

Acknowledgments

PL thanks Txema Heredia, Angel Carreno and Jordi Rambla for computational support. BP thanks Steven Gazal for his valuable assistance in writing scripts, Quentin Vincent for his advices on the datation protocol, and Marie-Claude Babron and Remi Kazma for their helpful comments on the manuscript.

Author Contributions

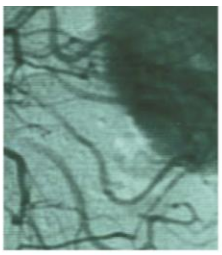
Conceived and designed the experiments: EG AS. Performed the experiments: BP PL HB. Analyzed the data: BP PL EG AS. Contributed reagents/materials/analysis tools: HB EP HMC. Wrote the paper: BP PL EP HMC EG AS.

References

1. Hirsh J, Dalen JE, Anderson DR, Poller L, Bussey H, et al. (1998) Oral anticoagulants: mechanism of action, clinical effectiveness, and optimal therapeutic range. Chest 114: 445S–469S.
2. Hyers TM, Agnelli G, Hull RD, Weg JG, Morris TA, et al. (1998) Antithrombotic therapy for venous thromboembolic disease. Chest 114: 561S–578S.
3. D'Andrea G, D'Ambrosio R, Margaglione M (2008) Oral anticoagulants: Pharmacogenetics Relationship between genetic and non-genetic factors. Blood Rev 22: 127–140.
4. Cooper GM, Johnson JA, Langaee TY, Feng H, Stanaway IB, et al. (2008) A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood 112: 1022–1027.
5. Takeuchi F, McGinnis R, Bourgeois S, Barnes C, Eriksson N, et al. (2009) A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet 5: e1000433.
6. Teichert M, Eijgelsheim M, Rivadeneira F, Uitterlinden AG, van Schaik RH, et al. (2009) A genome-wide association study of acenocoumarol maintenance dosage. Hum Mol Genet 18: 3758–3768.
7. Cha PC, Mushiroda T, Takahashi A, Kubo M, Minami S, et al. (2010) Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. Hum Mol Genet 19: 4735–4744.
8. Goldstein JA, de Morais SM (1994) Biochemistry and molecular biology of the human CYP2C subfamily. Pharmacogenetics 4: 285–299.
9. Stec DE, Roman RJ, Flasch A, Rieder MJ (2007) Functional polymorphism in human CYP4F2 decreases 20-HETE production. Physiol Genomics 30: 74–81.
10. Bardowell SA, Stec DE, Parker RS (2010) Common variants of cytochrome P450 4F2 exhibit altered vitamin E-{omega}-hydroxylase specific activity. J Nutr 140: 1901–1906.
11. Li T, Chang CY, Jin DY, Lin PJ, Khvorova A, et al. (2004) Identification of the gene for vitamin K epoxide reductase. Nature 427: 541–544.
12. Rost S, Fregin A, Ivaskevicius V, Conzelmann E, Hortnagel K, et al. (2004) Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature 427: 537–541.
13. Limdi NA, Wadelius M, Cavallari L, Eriksson N, Crawford DC, et al. (2010) Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups. Blood 115: 3827–3834.
14. Ross KA, Bigham AW, Edwards M, Gozdzik A, Suarez-Kurtz G, et al. (2010) Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. J Hum Genet 55: 582–589.
15. Li JZ, Absher DM, Tang H, Southwick AM, Casto AM, et al. (2008) Worldwide human relationships inferred from genome-wide patterns of variation. Science 319: 1100–1104.
16. Shriver MD, Kennedy GC, Parra EJ, Lawson HA, Sonpar V, et al. (2004) The genomic distribution of population substructure in four populations using 8,525 autosomal SNPs. Hum Genomics 1: 274–286.
17. Storz JF, Payseur BA, Nachman MW (2004) Genome scans of DNA variability in humans reveal evidence for selective sweeps outside of Africa. Mol Biol Evol 21: 1800–1811.
18. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585–595.
19. Frazer KA, Ballinger DG, Cox DR, Hinds DA, Stuve LL, et al. (2007) A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851–861.
20. Rieder MJ, Reiner AP, Gage BF, Nickerson DA, Eby CS, et al. (2005) Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med 352: 2285–2293.
21. Wu AH, Wang P, Smith A, Haller C, Drake K, et al. (2008) Dosing algorithm for warfarin using CYP2C9 and VKORC1 genotyping from a multi-ethnic population: comparison with other equations. Pharmacogenomics 9: 169–178.
22. Yang L, Ge W, Yu F, Zhu H (2010) Impact of VKORC1 gene polymorphism on interindividual and interethnic warfarin dosage requirement–a systematic review and meta analysis. Thromb Res 125: e159–166.
23. Smith JM, Haigh J (1974) The hitch-hiking effect of a favourable gene. Genet Res 23: 23–35.
24. Chen H, Patterson N, Reich D (2010) Population differentiation as a test for selective sweeps. Genome Res 20: 393–402.
25. Sabeti PC, Reich DE, Higgins JM, Levine HZ, Richter DJ, et al. (2002) Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832–837.
26. Weir BS, Hill WG (2002) Estimating F-statistics. Annu Rev Genet 36: 721–750.
27. Sabeti PC, Varilly P, Fry B, Lohmueller J, Hostetter E, et al. (2007) Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913–918.
28. Voight BF, Kudaravalli S, Wen X, Pritchard JK (2006) A map of recent positive selection in the human genome. PLoS Biol 4: e72.
29. Clark AG, Hubisz MJ, Bustamante CD, Williamson SH, Nielsen R (2005) Ascertainment bias in studies of human genome-wide polymorphism. Genome Res 15: 1496–1502.
30. Cann HM, de Toma C, Cazes L, Legrand MF, Morel V, et al. (2002) A human genome diversity cell line panel. Science 296: 261–262.
31. Genin E, Tullio-Pelet A, Begeot F, Lyonnet S, Abel L (2004) Estimating the age of rare disease mutations: the example of Triple-A syndrome. J Med Genet 41: 445–449.
32. Li J, Zhang L, Zhou H, Stoneking M, Tang K (2011) Global patterns of genetic diversity and signals of natural selection for human ADME genes. Hum Mol Genet 20: 528–540.
33. Marsh S, King CR, Porche-Sorbet RM, Scott-Horton TJ, Eby CS (2006) Population variation in VKORC1 haplotype structure. J Thromb Haemost 4: 473–474.
34. Limdi NA, Beasley TM, Crowley MR, Goldstein JA, Rieder MJ, et al. (2008) VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African-Americans and European-Americans. Pharmacogenomics 9: 1445–1458.
35. Schwarz UI, Ritchie MD, Bradford Y, Li C, Dudek SM, et al. (2008) Genetic determinants of response to warfarin during initial anticoagulation. N Engl J Med 358: 999–1008.
36. Geisen C, Watzka M, Sittinger K, Steffens M, Daugela L, et al. (2005) VKORC1 haplotypes and their impact on the inter-individual and inter-ethnical variability of oral anticoagulation. Thromb Haemost 94: 773–779.
37. Bodin L, Verstuyft C, Tregouet DA, Robert A, Dubert L, et al. (2005) Cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase (VKORC1) genotypes as determinants of acenocoumarol sensitivity. Blood 106: 135–140.
38. Carlson CS, Thomas DJ, Eberle MA, Swanson JE, Livingston RJ, et al. (2005) Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res 15: 1553–1565.
39. Wang ET, Kodama G, Baldi P, Moyzis RK (2006) Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc Natl Acad Sci U S A 103: 135–140.
40. Teo YY, Sim X, Ong RT, Tan AK, Chen J, et al. (2009) Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res 19: 2154–2162.
41. Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L (2008) Natural selection has driven population differentiation in modern humans. Nat Genet 40: 340–345.
42. Akey JM (2009) Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711–722.
43. Xue Y, Zhang X, Huang N, Daly A, Gillson CJ, et al. (2009) Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics 183: 1065–1077.
44. Hofer T, Ray N, Wegmann D, Excoffier L (2009) Large allele frequency differences between human continental groups are more likely to have occurred by drift during range expansions than by selection. Ann Hum Genet 73: 95–108.
45. Pickrell JK, Coop G, Novembre J, Kudaravalli S, Li JZ, et al. (2009) Signals of recent positive selection in a worldwide sample of human populations. Genome Res 19: 826–837.
46. Nath SK, Han S, Kim-Howard X, Kelly JA, Viswanathan P, et al. (2008) A nonsynonymous functional variant in integrin-alpha(M) (encoded by ITGAM) is associated with systemic lupus erythematosus. Nat Genet 40: 152–154.

47. Hom G, Graham RR, Modrek B, Taylor KE, Ortmann W, et al. (2008) Association of systemic lupus erythematosus with C8orf13-BLK and ITGAM-ITGAX. N Engl J Med 358: 900–909.
48. Jarvinen TM, Hellquist A, Koskenmies S, Einarsdottir E, Panelius J, et al. (2010) Polymorphisms of the ITGAM gene confer higher risk of discoid cutaneous than of systemic lupus erythematosus. PLoS One 5: e14212.
49. Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Akey JM (2011) Clusters of adaptive evolution in the human genome. Front Genet 2: 50.
50. Barreiro LB, Quintana-Murci L (2010) From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat Rev Genet 11: 17–30.
51. Suttie JW (1985) Vitamin K-dependent carboxylase. Annu Rev Biochem 54: 459–477.
52. Oldenburg J, Marinova M, Muller-Reible C, Watzka M (2008) The vitamin K cycle. Vitam Horm 78: 35–62.
53. Yan L, Zhou B, Greenberg D, Wang L, Nigdikar S, et al. (2004) Vitamin K status of older individuals in northern China is superior to that of older individuals in the UK. Br J Nutr 92: 939–945.
54. Beavan SR, Prentice A, Stirling DM, Dibba B, Yan L, et al. (2005) Ethnic differences in osteocalcin gamma-carboxylation, plasma phylloquinone (vitamin K1) and apolipoprotein E genotype. Eur J Clin Nutr 59: 72–81.
55. Kaneki M, Hodges SJ, Hosoi T, Fujiwara S, Lyons A, et al. (2001) Japanese fermented soybean food as the major determinant of the large geographic difference in circulating levels of vitamin K2: possible implications for hip-fracture risk. Nutrition 17: 315–321.
56. Fujita Y, Iki M, Tamaki J, Kouda K, Yura A, et al. (2011) Association between vitamin K intake from fermented soybeans, natto, and bone mineral density in elderly Japanese men: the Fujiwara-kyo Osteoporosis Risk in Men (FORMEN) study. Osteoporos Int.
57. Vermeer C, Shearer MJ, Zittermann A, Bolton-Smith C, Szulc P, et al. (2004) Beyond deficiency: potential benefits of increased intakes of vitamin K for bone and vascular health. Eur J Nutr 43: 325–335.
58. Wardrop D, Keeling D (2008) The story of the discovery of heparin and warfarin. Br J Haematol 141: 757–763.
59. Mueller RL, Scheidt S (1994) History of drugs for thrombotic disease. Discovery, development, and directions for the future. Circulation 89: 432–449.
60. Kohn MH, Pelz HJ, Wayne RK (2000) Natural selection mapping of the warfarin-resistance gene. Proc Natl Acad Sci U S A 97: 7911–7915.
61. Kohn MH, Pelz HJ, Wayne RK (2003) Locus-specific genetic differentiation at Rw among warfarin-resistant rat (Rattus norvegicus) populations. Genetics 164: 1055–1070.
62. Diaz JC, Song Y, Moore A, Borchert JN, Kohn MH (2010) Analysis of vkorc1 polymorphisms in Norway rats using the roof rat as outgroup. BMC Genet 11: 43.
63. Song Y, Endepols S, Klemann N, Richter D, Matuschka FR, et al. (2011) Adaptive introgression of anticoagulant rodent poison resistance by hybridization between old world mice. Curr Biol 21: 1296–1301.
64. Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, et al. (2011) Classic selective sweeps were rare in recent human evolution. Science 331: 920–924.
65. Pritchard JK, Di Rienzo A (2010) Adaptation - not by sweeps alone. Nat Rev Genet 11: 665–667.
66. Pritchard JK, Pickrell JK, Coop G (2010) The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr Biol 20: R208–215.
67. Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, et al. (2009) The role of geography in human adaptation. PLoS Genet 5: e1000500.
68. Consortium TGP (2010) A map of human genome variation from population-scale sequencing. Nature 467: 1061–1073.
69. Cann HM (1998) Human genome diversity. C R Acad Sci III 321: 443–446.
70. Rosenberg NA (2006) Standardized subsets of the HGDP-CEPH Human Genome Diversity Cell Line Panel, accounting for atypical and duplicated samples and pairs of close relatives. Ann Hum Genet 70: 841–847.
71. Stephens M, Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. Am J Hum Genet 73: 1162–1169.
72. Bandelt HJ, Forster P, Rohl A (1999) Median-joining networks for inferring intraspecific phylogenies. Mol Biol Evol 16: 37–48.
73. Barrett JC, Fry B, Maller J, Daly MJ (2005) Haploview: analysis and visualization of LD and haplotype maps. Bioinformatics 21: 263–265.
74. Luna A, Nicodemus KK (2007) snp.plotter: an R-based SNP/haplotype association and linkage disequilibrium plotting package. Bioinformatics 23: 774–776.
75. Lewontin RC (1964) The Interaction of Selection and Linkage. I. General Considerations; Heterotic Models. Genetics 49: 49–67.
76. Hill WG, Robertson A (1968) Linkage disequilibrium in finite populations Theoretical and Applied Genetics 38: 226–231.
77. Sabeti PC, Schaffner SF, Fry B, Lohmueller J, Varilly P, et al. (2006) Positive natural selection in the human lineage. Science 312: 1614–1620.
78. Wright S (1951) The genetical structure of populations. Annals of Eugenics 15 323–354.
79. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, et al. (2002) The Bioperl toolkit: Perl modules for the life sciences. Genome Res 12: 1611–1618.
80. Beaumont MA, Nichols RA (1996) Evaluating loci for use in the genetic analysis of population structure. The Royal Society 263: 1619–1626.
81. Gardner M, Bertranpetit J, Comas D (2008) Worldwide genetic variation in dopamine and serotonin pathway genes: implications for association studies. Am J Med Genet B Neuropsychiatr Genet 147B: 1070–1075.
82. Scheet P, Stephens M (2006) A fast and flexible statistical model for large-scale population genotype data: applications to inferring missing genotypes and haplotypic phase. Am J Hum Genet 78: 629–644.
83. McVean GA, Myers SR, Hunt S, Deloukas P, Bentley DR, et al. (2004) The fine-scale structure of recombination rate variation in the human genome. Science 304: 581–584.
84. Winckler W, Myers SR, Richter DJ, Onofrio RC, McDonald GJ, et al. (2005) Comparison of fine-scale recombination rates in humans and chimpanzees. Science 308: 107–111.

3. Brève Médecine/Science



© Inserm - Xavier Jeunemaitre

La variabilité de réponse aux anticoagulants oraux, une conséquence de la sélection naturelle

> La diversité génétique des populations humaines actuelles résulte de l'action de différences forces évolutives, comme la sélection naturelle, qui a joué un rôle déterminant dans la répartition des variants génétiques impliqués dans l'adaptation de l'homme à son environnement chimique. Certains de ces variants génétiques sont également impliqués dans la susceptibilité aux maladies ou dans la réponse aux médicaments. Des signatures de sélection naturelle peuvent ainsi être mises en évidence dans les régions génomiques qui contiennent des gènes impliqués dans ces traits, comme cela a pu être démontré dans une étude récente de la région génomique du gène *VKORC1* (vitamine K époxyde réductase) [1]. Ce gène code pour la cible pharmacologique des anticoagulants oraux de type antivitamine K (AVK), tels que la warfarine ou l'acénocoumarol, qui sont largement utilisés dans le traitement de la maladie thromboembolique veineuse ; il existe une très grande variabilité interindividuelle de réponse à ces molécules. Une part importante (30 %) de cette variabilité de réponse est expliquée par un polymorphisme (1639G > A) situé dans le promoteur du gène *VKORC1* [2]. L'étude, conduite dans les 52 populations du panel HGDP-CEPH (*human genome diversity cell line panel*) [3], montre que l'allèle 1639A, associé à une sensibilité accrue au traitement, a augmenté en fréquence jusqu'à atteindre quasiment une fixation sous l'action d'une sélection positive dans les 17 populations d'Asie de l'Est. La signature moléculaire de cette sélection n'est retrouvée dans aucune autre région du monde. L'événement sélectif, qui remonte à environ 4 500 ans, contribuerait ainsi aux différences de sensibilité génétique aux AVK observées aujourd'hui entre les populations humaines, les doses d'AVK requises étant en moyenne 1,5 à 2 fois moins élevées dans les populations asiatiques [4]. L'étude ne permet pas de déterminer si *VKORC1* est la cible directe de cette sélection puisque la région génomique identifiée contient également les gènes *MYST1* (*histone acetyltransferase 1*) et *PRSS8* (*protease serine 8*) impliqués dans la réponse immunitaire. Un argument plaide cependant en faveur d'une action directe de la sélection sur le gène *VKORC1* : on a observé, sur l'homologue murin de ce gène, l'apparition de nouvelles mutations dans les populations de rats d'égout en réponse à l'utilisation massive de la mort-au-rat, qui n'est autre qu'un anticoagulant de type AVK [5]. Un phénomène similaire pourrait ainsi être envisagé chez l'homme avec une adaptation des populations d'Asie de l'Est à une molécule naturelle de type AVK présente dans leur environnement chimique. ◊

Blandine Patillon
Emmanuelle Génin
Audrey Sabbagh
Inserm UMR 946, IRD UMR 216, Paris, France
audrey.sabbagh@parisdescartes.fr

1. Patillon B, *et al. PLoS One* 2013 ; 7 : e53049.
2. D'Andrea G, *et al. Blood Rev* 2008 ; 22 : 127-40.
3. Cann HM, *et al. Science* 2002 ; 296 : 261-2.
4. Limdi NA, *et al. Blood* 2010 ; 115 : 3827-34.
5. Rost S, *et al. BMC Genet* 2009 ; 10 : 4.

Chapitre 3

Recherche de la cible de sélection dans la région génomique de *VKORC1* : apport des données du Projet 1000 Génomes

Nous avons vu dans le chapitre précédent que nous ne pouvions pas, à partir des données de génotypage dont nous disposions, identifier de manière précise la cible de la pression de sélection qui est intervenue dans la région génomique de *VKORC1* en Asie de l'Est. Pour déterminer dans quelle mesure notre capacité de localisation était limitée par la densité des données génétiques utilisées, nous avons cherché à voir si nous pouvions préciser le signal de sélection dans la région en utilisant des données de séquençage. Nous avons donc mis à profit la disponibilité récente des données du Projet 1000 Génomes (1KG) en procédant à une analyse approfondie des signatures génomiques du balayage sélectif précédemment détecté au locus de *VKORC1* dans les données du Panel HGDP-CEPH. Dans ce chapitre sont présentés les résultats de ce travail.

1. Matériel et Méthodes

Nous avons extrait des données 1KG les variants de types SNVs présents dans une région de 2 Mégabases centrée sur *VKORC1*. Comme nous l'avons expliqué plus en détail dans la partie 3 de cette thèse, nous n'avons retenu que les variants qui étaient présents dans les données de faible couverture (*low coverage whole genome data*) et exclu les polymorphismes d'insertion-délétion et les variations du nombre de copies.

Nous avons appliqué sur ces données trois des tests de sélection précédemment utilisés sur les données de génotypage HGDP-CEPH, à savoir les tests XP-CLR, XP-EHH et F_{ST}, les plus puissants pour détecter des phénomènes de balayage sélectif complet ou presque complet, conduisant à la quasi-fixation de l'allèle sélectionné, comme cela semble être le cas dans cette région du génome en Asie de l'Est. Nous avons également utilisé un quatrième test, la statistique *D* de Tajima, qui nécessite de disposer de données non biaisées sur le nombre de sites polymorphes dans une région génomique. En conséquence il n'est pas possible d'utiliser ce test de sélection sur des données de génotypage, mais uniquement sur des données de séquences.

2. Étude du signal de sélection

Recherche du signal de sélection au niveau mondial

Dans un premier temps, nous avons voulu vérifier que le signal de sélection étudié sur les données HGDP-CEPH était bien détecté en Asie de l'Est et uniquement dans cette région géographique dans les données 1KG. Nous avons appliqué les différents tests de sélection dans trois populations situées sur les trois continents Afrique, Europe et Asie : les Yoruba du Nigéria (YRI), les résidents de l'Utah d'origine européenne (CEU), et les Chinois Han (CHB). Pour cette étude, nous n'avons pas considéré les populations d'Amérique, qui sont fortement mélangées (1000 Genomes Project Consortium et al., 2012), ce qui impacte le spectre de fréquence allélique des variants et risque de biaiser les résultats des tests de sélection qui en tiennent compte. Les tests XP-CLR et XP-EHH requérant la définition d'une population de référence, nous avons choisi les Yoruba comme référence pour les populations d'origine non africaine (CEU et CHB), et les CEU pour les Yoruba.

Pour tous les tests employés, un signal de sélection est bien détecté au niveau du locus de *VKORC1* au seuil de significativité 0,05, chez les Chinois Han et uniquement dans cette population (Figure 2.6). Les tests F_{ST}, XP-CLR et XP-EHH présentent des scores significatifs au seuil de 1 %. Dans cette population, un pic de valeurs significatives ($P < 0,01$) est également

observable en aval avec le *D* de Tajima et les scores XP-EHH. Ce signal de sélection correspond en réalité au « troisième pic d'XP-EHH » dont nous parlons dans l'article 1. Il se situe au niveau des gènes *ITGAM* et *ITGAX* et est indépendant du signal sélectif concernant le locus *VKORC1* (absence de déséquilibre de liaison entre ces signaux).

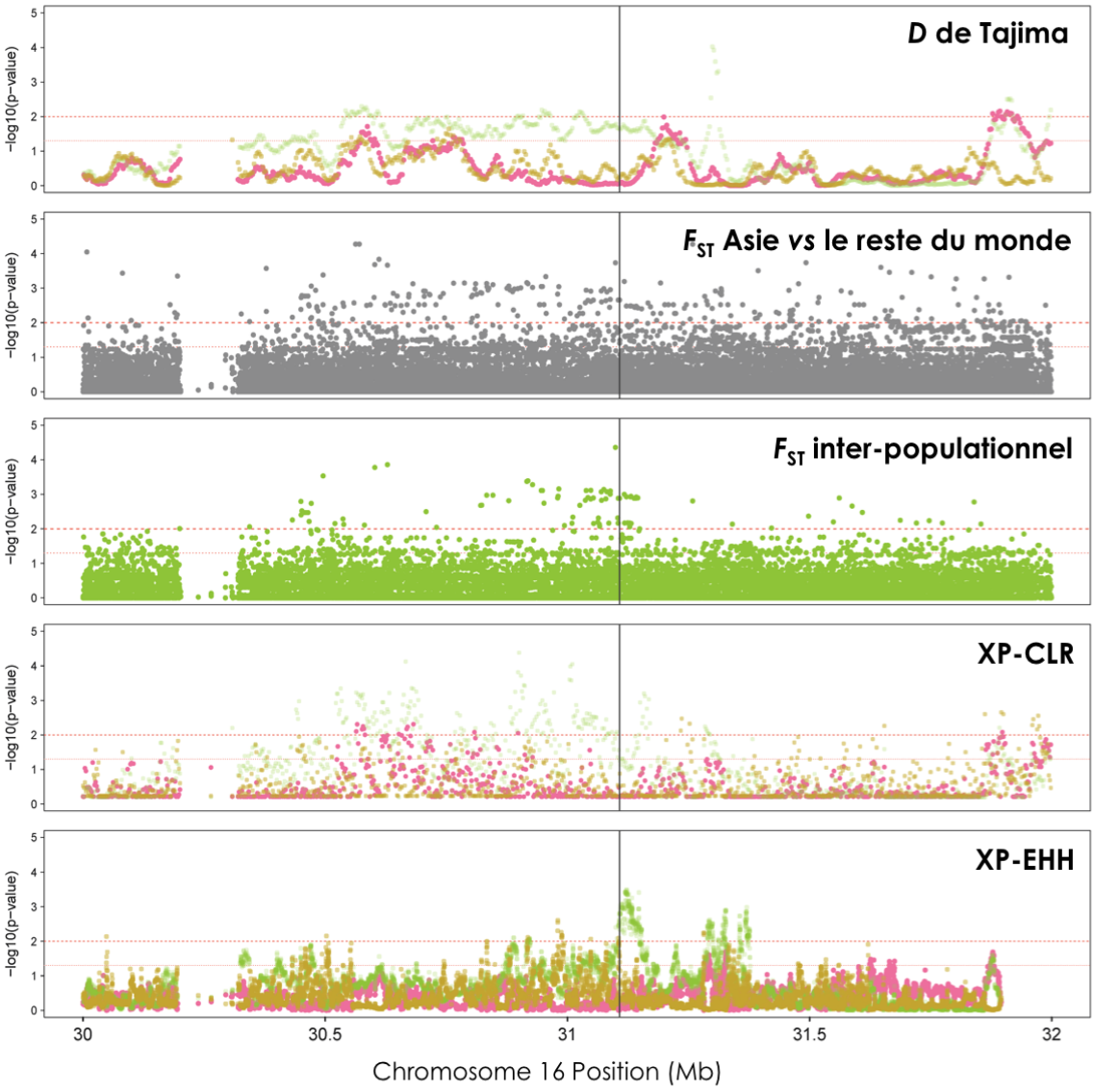


Figure 2.6 | Distribution des *P*-values des tests *D* de Tajima, F_{ST}, XP-CLR et XP-EHH dans une région de 2 Mb centrée sur *VKORC1*. Les trois statistiques *D* de Tajima, XP-CLR et XP-EHH ont été appliquées dans les populations CHB, CEU et YRI, présentées en vert, rose et orange, respectivement. Pour l'XP-CLR et l'XP-EHH, la population de référence pour les tests appliqués dans les populations CHB et CEU est la population YRI ; la population CEU pour les tests appliqués dans la population YRI. Le F_{ST} a été estimé à deux niveaux : entre les populations (gris) et entre l'Asie de l'Est et les autres régions géographiques (vert).

Recherche du signal de sélection dans toutes les populations d'Asie de l'Est

Afin de vérifier que le balayage sélectif mis en évidence dans la population CHB était également détecté dans les autres populations d'Asie de l'Est échantillonnées dans le Projet 1000 Génomes, nous avons recherché la présence du signal de sélection dans la population de Chinois Han vivant dans le sud de la Chine (CHS) et la population de Japonais (JPT). Pour cela, nous nous sommes concentrés sur les tests F_{ST} et XP-EHH qui donnaient les signaux les plus nets dans la population CHB.

Nous avons bien détecté la signature génomique de la sélection positive dans ces deux populations, comme le montre la Figure 2.7. Les profils de distribution des *P*-values entre les trois populations asiatiques sont extrêmement proches, si bien qu'il est suffisant de se concentrer sur l'étude d'une seule population. Par conséquent nous ne présenterons par la suite que les résultats obtenus dans la population CHB.

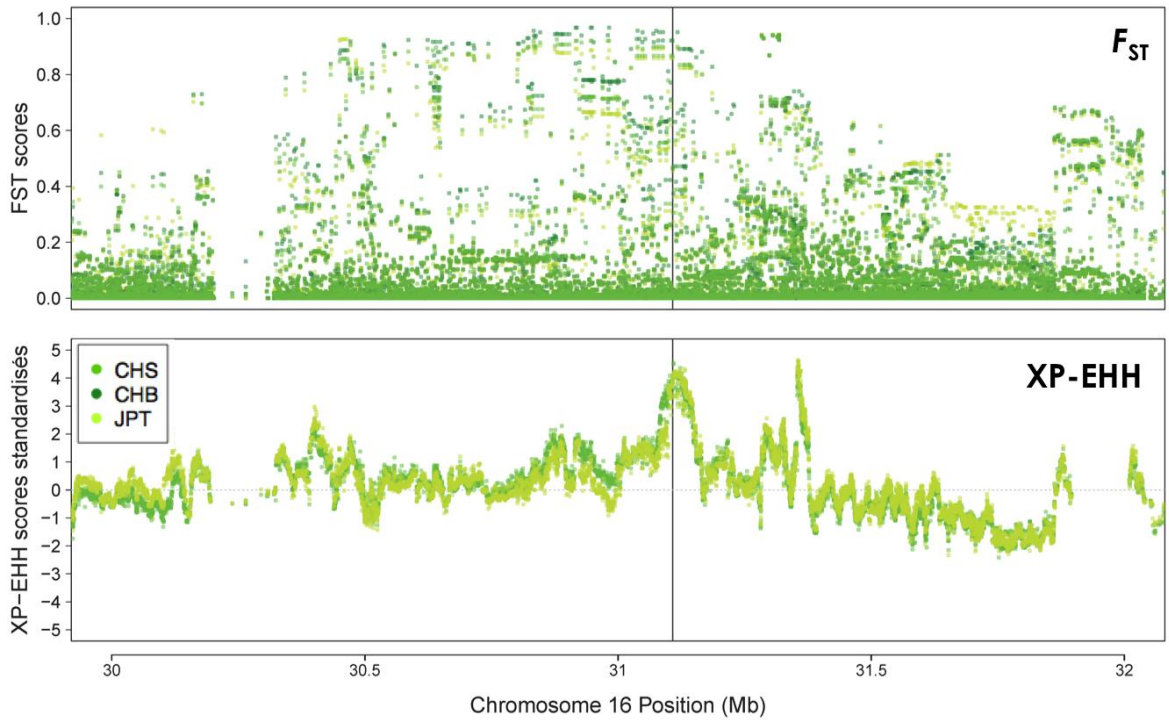


Figure 2.7 | Distribution des *score* F_{ST} et XP-EHH appliqués dans une région de 2 Mb centrée sur *VKORC1* dans les trois populations asiatiques (CHB, CHS et JPT). Le F_{ST} a été calculé en comparant chacune des populations asiatiques avec la population des Yoruba d'Afrique (YRI). L'XP-EHH a été appliqué en utilisant la population YRI.

Délimitation du signal de sélection

Utilisant la même stratégie que lors de l'étude précédente sur les données HGDP-CEPH, nous nous sommes servis des clusters de valeurs significatives ($P < 0,01$) obtenus dans la population CHB, pour essayer de délimiter précisément la région génomique sur laquelle s'étend le signal de sélection.

Pour délimiter le signal de sélection, nous nous sommes focalisés sur la région génomique de 499 kb (30,66 - 31,16), correspondant au bloc de déséquilibre de liaison identifié en Asie de l'Est dans les données HDGP-CEPH, allant du gène *PRR14* au gène *PRSS8*. Nous avons inclus le gène *PRSS36,* voisin de *PRSS8*, qui n'était pas couvert par les données HGDP-CPEH, et qui appartient à ce bloc de déséquilibre de liaison. Au total, 26 gènes sont présents dans ce bloc de déséquilibre de liaison.

Recherche du gène cible de la sélection

Nous avons regardé la distribution du F_{ST} inter-populationnel (Figure 2.8), F_{ST} comparant l'Asie de l'Est au reste du monde (Figure 2.9) et XP-EHH (Figure 2.9) dans cette région génomique de 499 kb.

Des valeurs élevées de F_{ST} sont observées pour de nombreux variants dans cette région, à la fois lorsque l'on compare les 14 populations et lorsque l'on compare l'Asie de l'Est au reste du monde. Un variant se détache cependant assez nettement des autres : il s'agit du variant rs11150606 dans le gène *PRSS53*.

Quant au test XP-EHH, les valeurs les plus extrêmes (> 4) sont restreintes aux quatre gènes adjacents : *VKORC1 – BCKDK – KAT8 – PRSS8* (Figure 2.10).

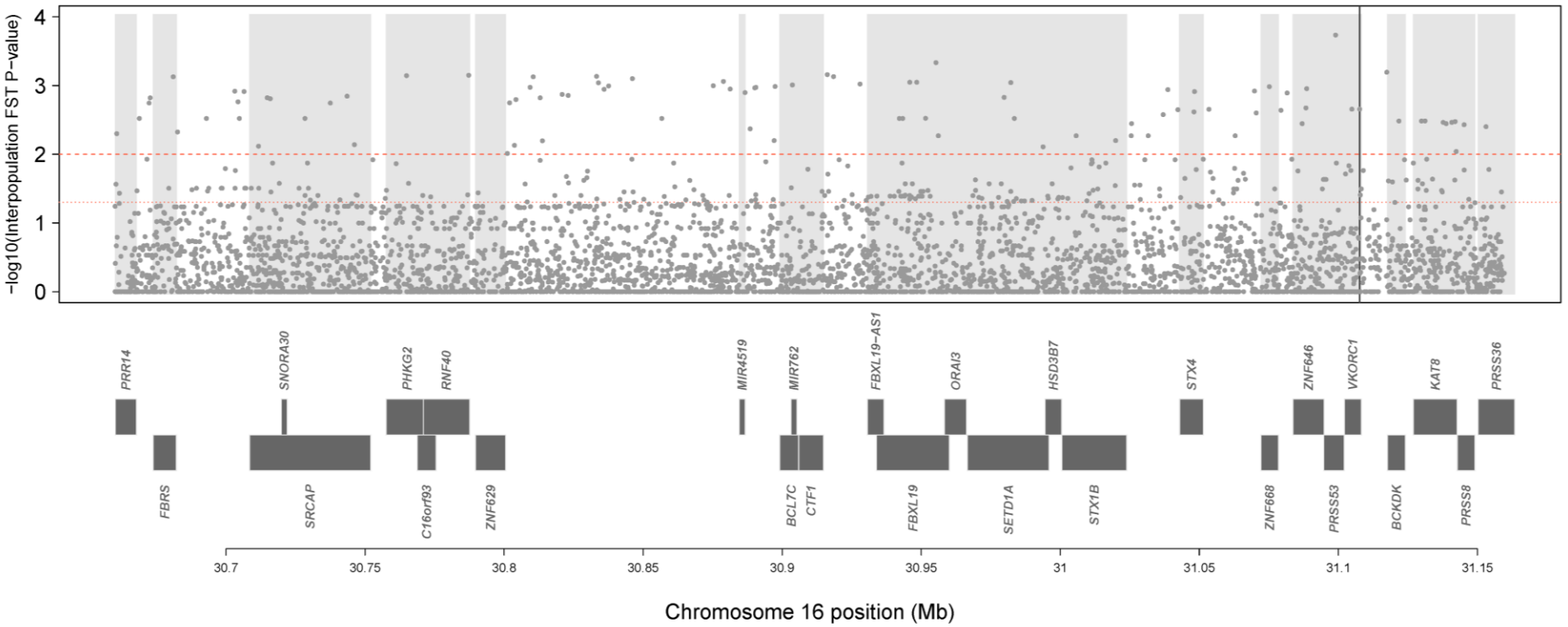


Figure 2.8 | Distribution des *P*-values du F_{ST} inter-populationnel au sein de la région génomique de 499 kb contenant *VKORC1*. Le F_{ST} a été calculé entre les 14 populations échantillonnées dans 1000 Génomes.

La combinaison de ces résultats nous a permis de sélectionner cinq gènes adjacents : *PRSS53 – VKORC1 – BCKDK – KAT8* et *PRSS8*, qui en conséquence représentent de bons candidats pour héberger le variant génétique ciblé par la sélection. Par rapport à la liste de gènes candidats établie sur les données HGDP-CEPH, nous avons donc identifié un gène supplémentaire (*PRSS53*). Il est utile de noter que ce gène est nommé *POL3S* dans l'article 1.
Ce gène, qui code pour une protéase à sérine, a été découvert récemment (Cal et al., 2006), et est très peu étudié. Des associations ont été retrouvées entre des variants de ce gène et le psoriasis (Stuart et al., 2010) et la maladie de Parkinson (Pankratz et al., 2012).

Il est intéressant de remarquer que le variant rs11150606 de *PRSS53* est celui qui, parmi l'ensemble des variants de ces cinq gènes candidats, possède la valeur de F_{ST} maximale, que ce soit pour le F_{ST} estimé au niveau inter-populationnel (F_{ST} = 0,579) ou entre l'Asie de l'Est et le reste du monde (F_{ST} = 0,748). Ce variant est une substitution non synonyme (T>C Gln30Arg) et représente donc un bon candidat pour être la cible directe de la sélection naturelle.

Apport des données de séquence

Il est clair que l'utilisation de données de séquençage permet d'améliorer considérablement la densité de l'investigation du génome, et d'obtenir ainsi une vision beaucoup plus détaillée de la signature génomique laissée par la sélection naturelle.

Sur la Figure 2.9 ci-dessous, nous voyons clairement que la densité des scores obtenus dans les données 1KG est telle qu'elle nous permet d'étudier un nombre de variants bien plus élevé, et surtout, d'explorer des gènes supplémentaires sur des locus du chromosome 16 pour lesquels nous ne disposions d'aucun variant génotypé dans les données HGDP-CEPH, comme les gènes *PRSS53* et *PRSS36*.

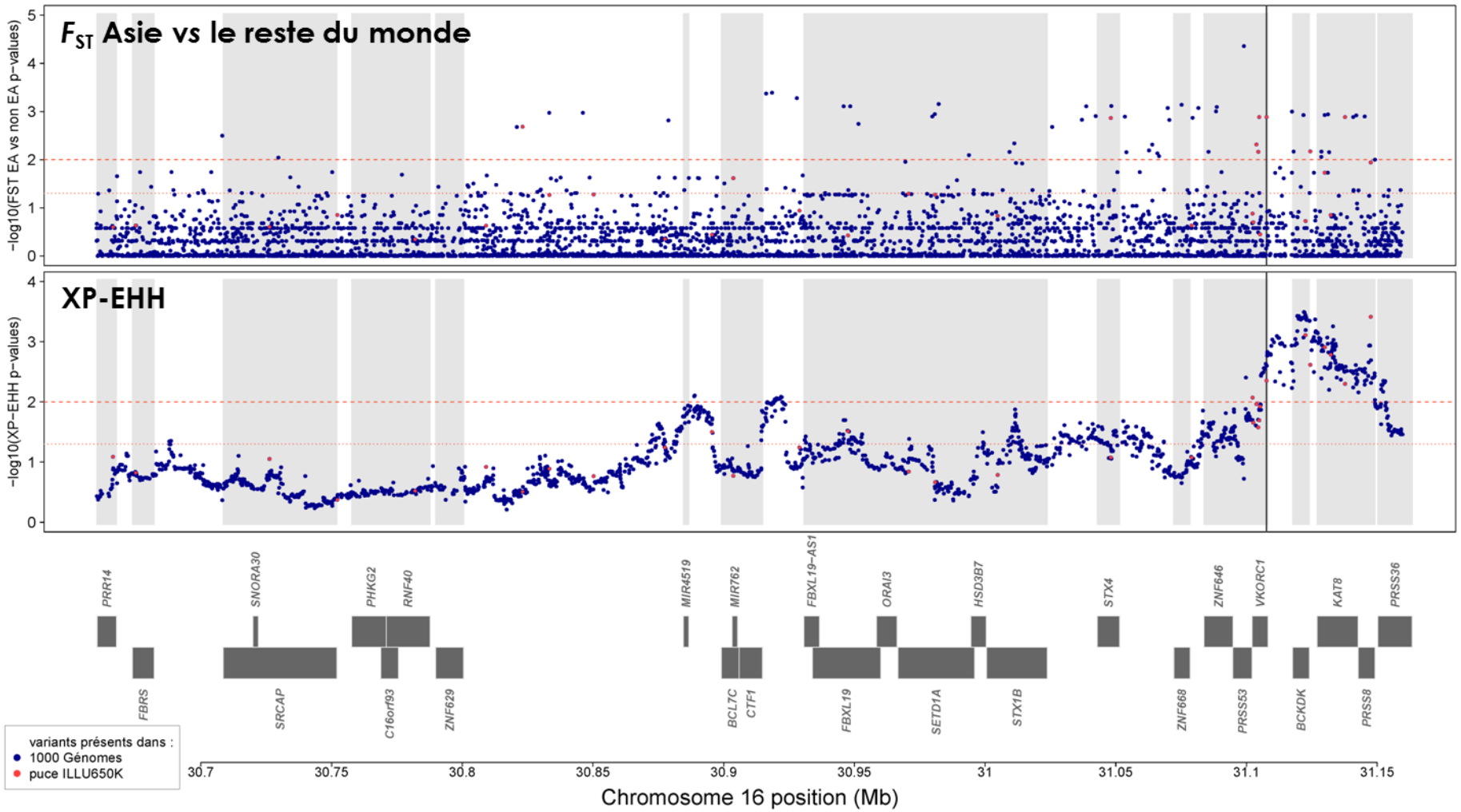


Figure 2.9| Résultats des *P*-values des tests F_{ST} comparant l'Asie de l'Est aux autres régions géographiques et XP-EHH dans les données 1000 Génomes au sein de la région génomique de 499 kb contenant *VKORC1*. Le F_{ST} a été calculé en comparant les individus des trois populations asiatiques (CHB, CHS et JPT) *versus* les individus des 11 autres populations mises ensemble. La statistique XP-EHH a été appliquée dans la population de Chinois Han (CHB) en prenant la population africaine des Yoruba (YRI) comme population de référence. L'ensemble des variants représentés correspond aux variants annotés dans les données de 1KG. En rouge sont mis en évidence les variants qui étaient présents sur la puce ILLU650K utilisée pour le génotypage du panel HGDP-CEPH.

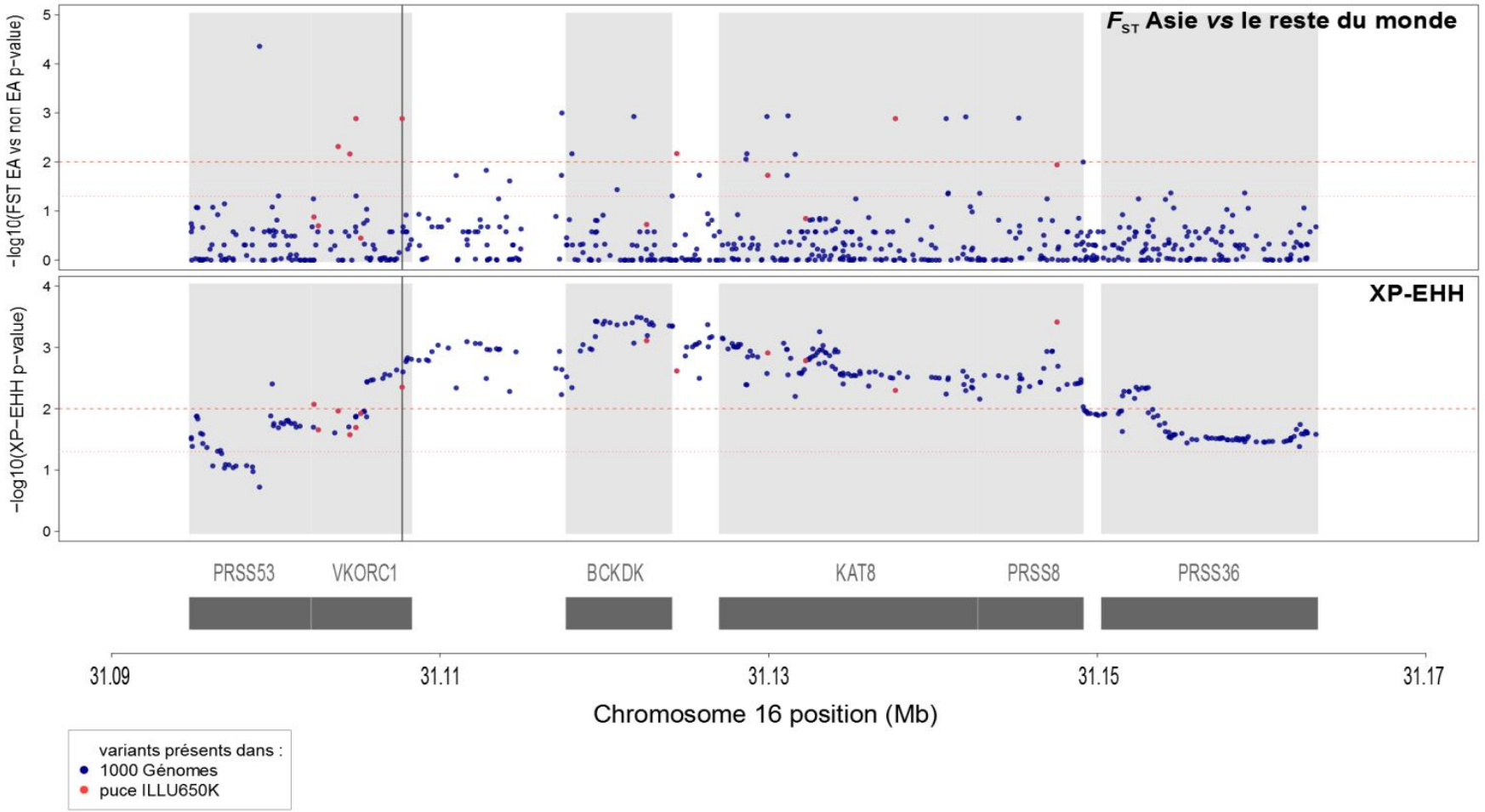


Figure 2.10 | Résultats des *P*-values des tests F_{ST} comparant l'Asie de l'Est aux autres régions géographiques et XP-EHH dans les données 1000 Génomes au locus de *VKORC1*. Le F_{ST} a été calculé en comparant les individus des trois populations asiatiques (CHB, CHS et JPT) *versus* les individus des 11 autres populations mises ensemble. La statistique XP-EHH a été appliquée dans la population de Chinois Han (CHB) en prenant la population africaine des Yoruba (YRI) comme population de référence. L'ensemble des variants représentés correspond aux variants annotés dans les données de 1KG. En rouge sont mis en évidence les variants qui étaient présents sur la puce ILLU650K utilisée pour le génotypage du panel HGDP-CEPH.

Par ailleurs, la forme de la distribution des scores des tests XP-EHH et XP-CLR dans la région des 2 Mb a changé par rapport à ce que l'on observait dans les données HGDP-CEPH, comme le montre la Figure 2.11. Alors que trois clusters distincts de scores extrêmes ($P < 0,01$) étaient détectables avec ces deux tests dans les données HGDP-CEPH, seulement deux sont encore retrouvés dans les données 1KG avec l'XP-EHH. Il est intéressant de remarquer que le signal le plus fort détecté à présent avec l'XP-EHH concerne notre signature d'intérêt dans la région génomique autour du gène *VKORC1*. Quant à l'XP-CLR, la distribution de ces scores est plus diffuse dans les données 1KG, ne permettant pas d'isoler de façon non ambigüe un pic de scores extrêmes dans le signal englobant ces trois clusters.

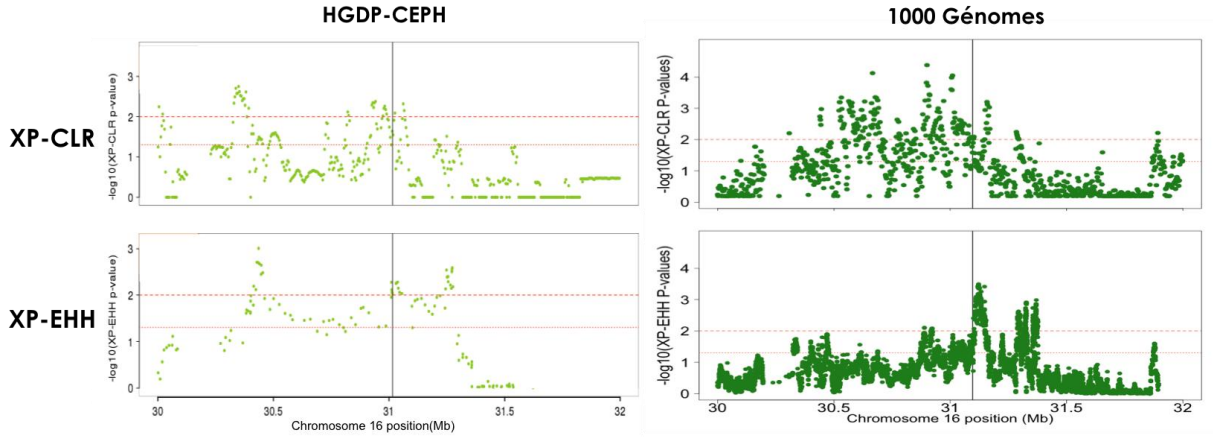


Figure 2.11 | Distribution des *P*-values des tests XP-CLR et XP-EHH dans une région de 2 Mb centrée sur *VKORC1* dans les données HGDP-CEPH et 1KG. Pour les données HGDP-CEPH, les tests ont été appliqués dans la région d'Asie de l'Est et pour les données 1KG, dans la population CHB. La position du variant fonctionnel de *VKORC1* rs9923231 est indiquée par une barre verticale grise.

Étude du gène PRSS53

Nos résultats précédents ayant montré que le SNP rs11150606 du gène *PRSS53* présentait un profil de différentiation particulièrement élevé entre l'Asie de l'Est et le reste du monde, nous nous sommes intéressés plus particulièrement à ce variant. Comme le montre la Figure 2.12, nous pouvons constater qu'à

l'instar de l'allèle dérivé T[8] du variant fonctionnel rs9923231 de *VKORC1*, l'allèle dérivé C de ce variant rs11150606 est majoritaire dans les populations asiatiques (79 %) alors qu'il est quasiment absent dans les populations africaines. Il est rare en Europe (4 %) et plus fréquent dans les populations d'Amérique (18 %).

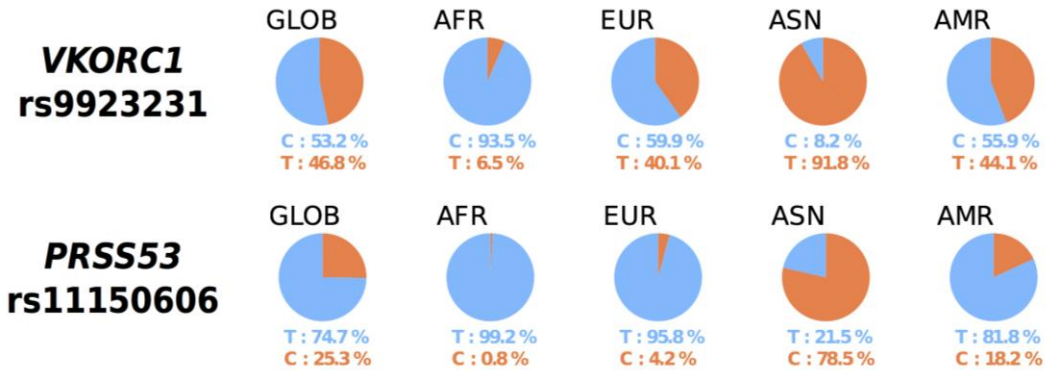


Figure 2.12 | Distribution des fréquences alléliques des SNPs rs9923231 de *VKORC1* et rs11150606 de *PRSS53* au niveau global et dans les quatre régions géographiques de 1000 Génomes. Les allèles dérivé et ancestral sont représentés en orange et bleu, respectivement.

Afin de regarder le lien entre le variant rs11150606 de *PRSS53* et le variant fonctionnel rs9923231 de *VKORC1*, nous avons regardé quels haplotypes portaient l'allèle dérivé de ces deux variants. Nous avons reconstruit les haplotypes à trois niveaux : i) au niveau global (en considérant les 1 089 individus), ii) dans toutes les régions hors Asie de l'Est (en considérant les 803 individus d'Afrique, d'Europe et d'Amérique) et iii) en Asie de l'Est (en considérant uniquement les 286 individus des trois populations de ce continent), à partir de l'ensemble des variants de ces deux gènes présents à une MAF > 0,01 au niveau global. Les haplotypes reconstruits ainsi que leurs fréquences sont présentés dans la Figure 2.13.

On remarque qu'il n'existe qu'un seul haplotype portant l'allèle dérivé de ces deux variants (H1). Cet haplotype est retrouvé à des fréquences très variables au niveau global (0,252), hors Asie de l'Est (0,063) et en Asie de l'Est (0,785)

[8] Il est à noter ici que dans les données 1000 Génomes, les allèles de ce SNP sont T et C, et non pas A et G comme dans les données HGDP-CEPH, ce qui est lié à une différence dans les brins d'ADN considérés dans ces deux études (brin + ou brin -).

(Figure 2.13). En Asie de l'Est, deux haplotypes portant l'allèle dérivé T du variant fonctionnel rs9923231 de *VKORC1* sont observés : H1 et H3. Un troisième haplotype portant cet allèle (H2) est observé à des fréquences relativement faibles au niveau global et hors Asie (respectivement 0,128 et 0,173).

De manière très intéressante, on s'aperçoit que la fréquence en Asie de l'haplotype H3 portant le variant dérivé T de rs9923231 et l'allèle ancestral T de rs11150606 est bien plus faible (0,131) que celle de l'haplotype H1 portant les deux allèles dérivés. Autrement dit, il semble que dans cette région géographique, la fréquence élevée de l'allèle T de ce variant soit conditionnée par celle de l'allèle dérivé du variant rs11150606 de *PRSS53*. Ce résultat suggère que le variant rs11150606 de *PRSS53* est probablement apparu plus récemment que le variant fonctionnel rs9923231 de *VKORC1* en Asie de l'Est.

Ces variants sont en déséquilibre de liaison élevé en Asie de l'Est, mais les fréquences alléliques de ces variants ne sont pas les mêmes (D' = 1, r^2 = 0,384), c'est-à-dire que si l'allèle dérivé C du variant rs11150606 est toujours retrouvé avec l'allèle dérivé T du variant rs9923231, l'inverse n'est pas vrai.

Haplotype	Séquence haplotypique																				Fréquence		
																					Globale	Hors Asie de l'Est[1]	Asie de l'Est[2]
H1	G	G	C	G	G	C	G	A	C	T	G	C	G	T	A	G	A	G	A	T	T	0,252	0,063	0,785
H2	G	G	C	G	G	C	G	A	T	T	G	C	G	T	A	G	A	G	C	T	T	0,128	0,173	0,000
H3	G	G	C	G	G	C	G	A	T	T	G	C	G	T	A	G	A	G	A	T	T	0,086	0,070	0,131
H4	G	G	T	C	C	C	C	C	T	C	A	T	G	T	G	C	G	G	A	C	T	0,192	0,234	0,082
H5	A	G	C	G	G	C	G	A	T	T	G	C	G	T	G	C	G	A	A	C	T	0,117	0,159	0,000
H6	G	A	T	C	C	T	C	C	T	C	G	T	A	T	G	C	G	G	A	C	C	0,057	0,077	0,000
H7	G	G	T	C	C	C	C	C	T	C	A	T	G	C	G	C	G	G	A	C	T	0,055	0,072	0,000
H8	G	G	T	C	C	C	C	C	T	C	G	C	G	T	G	G	G	G	A	C	C	0,041	0,056	0,000
H9	G	G	C	G	G	C	G	A	T	T	G	C	G	T	A	C	G	G	A	C	T	0,030	0,041	0,000
H10	G	G	C	G	G	C	G	A	T	T	G	C	G	T	G	C	G	G	A	C	T	0,019	0,026	0,000
H11	G	G	C	G	G	C	G	C	T	T	G	C	G	T	G	C	G	G	A	C	T	0,013	0,017	0,000
	rs45599534	rs751117	rs11865038	rs11864806	rs11864839	rs56909237	rs7199949	rs4468641	**rs11150606**	rs17839568	rs73530203	rs7294	rs7200749	rs17882023	rs2359612	rs8050894	rs9934438	rs17708472	rs2884737	**rs9923231**	rs17878544			
	PRSS53											***VKORC1***											

[1] Populations d'Afrique (ASW, LWK et YRI), d'Europe (CEU, FIN, GBR, IBS et TSI) et d'Amérique (CLM, MXL et PUR) mises ensemble.

[2] Populations d'Asie (CHB, CHS, et JPT) mises ensemble.

Figure 2.13 | Distribution des haplotypes de *PRSS53* et *VKORC1* au niveau global, hors Asie de l'Est et en Asie de l'Est. Pour chaque haplotype, les variants sont listés dans l'ordre de la position physique. Les allèles ancestral et dérivé sont représentés respectivement en bleu et orange. Les haplotypes sont triés par i) présence de l'allèle dérivé du variant rs11150606 de *PRSS53*, puis par ii) présence de l'allèle dérivé du variant rs9923231 de *VKORC1*, et enfin par iii) fréquence globale décroissante. Les haplotypes ayant une fréquence < 1 % au niveau global ne sont pas représentés. La reconstruction haplotypique a été réalisée avec l'algorithme EM implémenté dans le logiciel Haploview (Barrett et al., 2005), à partir de l'ensemble des variants ayant une MAF globale > 1 %.

3. Discussion et conclusion

Nous confirmons la présence d'un fort signal de sélection positive au locus du chromosome 16p11.2, en Asie de l'Est. Dans cette région du génome très riche en gènes, nous sommes finalement parvenus à identifier un nouveau variant candidat pour être la cible de la sélection positive dans les données de séquençage de 1KG. Ce variant non synonyme rs11150606 de *PRSS53* n'était pas présent, ni aucun variant dans ce gène, dans les données de la puce Illumina 650K dont nous disposions pour notre étude précédente (article 1).

Il est intéressant de remarquer que l'application aux données de séquences 1KG du *D* de Tajima, seul test que nous ne pouvions pas appliquer sur les données de génotypage d'HGDP-CEPH du fait du biais de sélection des variants qui affectait ces données, ne nous a finalement pas été utile pour affiner la localisation du signal sélectif identifié. En effet ce test affichait des scores significatifs ($P < 0,05$) dans la région avoisinant *VKORC1* en Asie de l'Est et uniquement dans ce continent (Figure 2.6), mais ne montrait pas de regroupement de scores très significatifs qui pourraient pointer une région réduite du génome qui serait plus susceptible d'héberger la cible de sélection.

Il est probable que le signal révélé par le F_{ST} sur le variant candidat rs11150606 soit confirmé par le test de sélection CMS, qui combine plusieurs statistiques de tests de sélection et possède une puissance de détection et de localisation de la cible de sélection supérieure à celles des tests individuels que nous avons utilisés (*cf.* chapitre 2 de la partie 1) (Grossman et al., 2010). Afin de le vérifier, nous avons regardé la distribution des scores calculés sur les données de la phase Pilote du Projet 1KG en Asie de l'Est dans la région génomique de 2 Mb autour de *VKORC1* (Figure 2.14). Nous avons observé une valeur extrême du score CMS pour le variant rs11150606 de *PRSS53*, nettement supérieure à celles calculées pour les variants voisins, confirmant nos résultats. Cependant, ces données ne comprenaient pas les données de couverture dense de l'ensemble du génome, et il est tout à fait possible,

qu'une fois ce test appliqué sur les données 1KG actualisées, des variants proches de rs11150606 puissent également ressortir de façon très significative. L'application du test CMS aux données complètes de 1KG nous aurait probablement permis de répondre à cette question, mais malheureusement nous n'avons pas réussi à utiliser le programme mis à disposition par les auteurs.

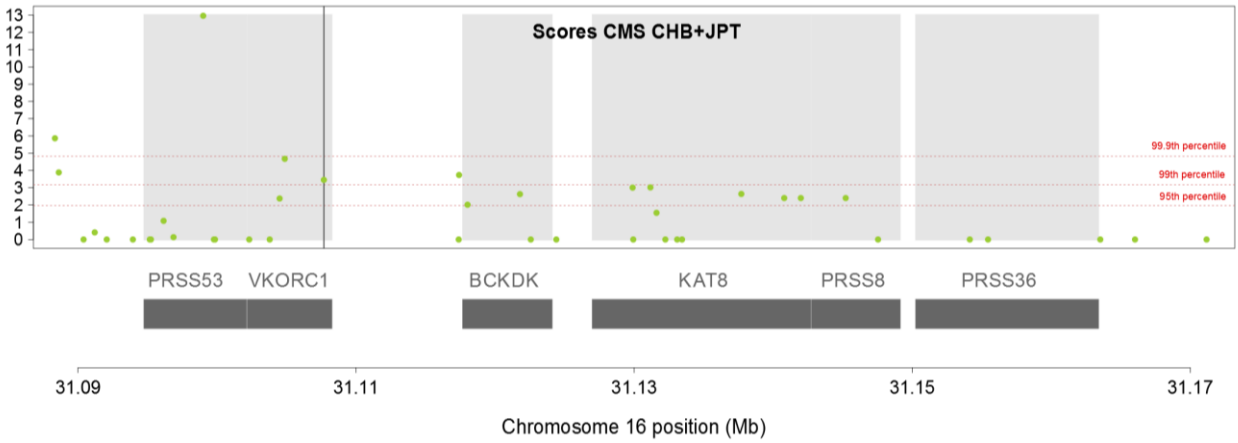


Figure 2.14 | Distribution des scores du CMS dans une région de 2 Mb centrée sur *VKORC1*. Les scores calculés dans les populations JPT+CHB comme indiqué dans Grossman *et al.* (2013) ont été téléchargés à partir du CMS browser (http://www.broadinstitute.org/scientific-community/science/programs/medical-and-population-genetics/cms/results). Les percentiles ont été estimés à partir de la distribution génome entier de ces scores.

Notre analyse ne comprend que deux des statistiques qui font partie du calcul du score CMS (F_{ST} et XP-EHH). Parmi les trois autres incluses dans ce test, l'iHS et le ΔiHH, basées sur la mesure de l'EHH (comme l'XP-EHH), ne nous auraient sans doute pas apporté grand-chose, car aucune d'elles ne bénéficient d'une bonne puissance de détection de la sélection lorsqu'il s'agit d'un balayage sélectif presque complet qui a entraîné les allèles dérivés à des fréquences proches de la fixation, comme c'est le cas ici. En revanche, la statistique ΔDAF, qui compare la fréquence de l'allèle dérivé dans la population testée à la fréquence moyenne de l'allèle dérivé dans d'autres populations, aurait sans doute été plus puissante. Étant donné que l'allèle dérivé du variant rs11150606 de *PRSS53* est retrouvé moins fréquemment dans les populations européennes et américaines que celui du variant fonctionnel rs9923231 de *VKORC1*, il est probable que ce score aurait

été supérieur pour le variant de *PRSS53*. Par manque de temps, nous n'avons malheureusement pas pu calculer cette statistique.

Néanmoins, sur les cinq statistiques utilisées par le CMS, les deux que nous avons appliquées (F_{ST} et XP-EHH) correspondent aux deux possédant la meilleure puissance de résolution spatiale de la cible de sélection (suivies par le ΔDAF) (Grossman et al., 2010). L'apport des autres statistiques n'aurait donc sans doute pas contribué de façon majeure à la localisation du variant causal. Or, la combinaison de ces seuls deux tests ne nous permet pas de pointer de variant aussi nettement que le fait le CMS sur les données pilotes de 1KG, puisque si le F_{ST} pointe largement ce variant, l'XP-EHH n'est pas significatif pour ce variant (P = 0,189).

En définitive, il apparaît que l'identification claire du gène ciblé par la sélection, et plus précisément du variant causal, relève dans notre cas de l'utopie… Pourtant, il est important de constater que notre difficulté à identifier la cible de sélection, rencontrée à deux reprises, soit amplifiée par deux aspects caractérisant le balayage sélectif étudié : il s'est produit en premier lieu, au niveau d'une région génomique très riche en gènes ; et en second, dans des populations d'Asie de l'Est, caractérisées par un niveau de déséquilibre de liaison très élevé. Si l'évènement sélectif était apparu dans des populations africaines, dans lesquelles le déséquilibre de liaison est moindre ; le variant ciblé par la sélection aurait peut-être été plus facilement identifiable.

Partie 3

Différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments

Chapitre 1

Contexte

Il y a 4 500 ans en Asie de l'Est, la sélection naturelle a été à l'origine d'une augmentation importante de la fréquence de l'allèle dérivé du polymorphisme fonctionnel rs9923231 de *VKORC1* conférant un phénotype de sensibilité augmentée aux AVK dans ces populations. Cet exemple illustre bien l'intérêt qu'il y a à connaître la structure génétique des populations humaines pour les gènes de la réponse aux médicaments, afin de tenir compte de ces différences dans l'optimisation des thérapies médicamenteuses. Il révèle également le rôle déterminant de la sélection naturelle dans la formation de profils de différenciation géographique atypiques pour des gènes d'intérêt majeur dans les phénotypes de la réponse aux médicaments. En effet, du fait de leur rôle de médiateurs entre l'organisme et l'environnement, les enzymes intervenant dans le métabolisme, le transport, et l'élimination des médicaments, ainsi que les récepteurs et cibles des médicaments représentent des candidats de choix pour l'action de la sélection naturelle.

1. Bilan des études analysant la différenciation pharmacogénétique des populations humaines

Comme nous l'avons mentionné dans la partie 1, de nombreuses études se sont intéressées à la différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments (McCarthy et al., 2002).

Importante diversité pharmacogénétique inter-populationnelle

Quels que soient le nombre de gènes et de populations considérés, ces études ont quasiment toutes révélé une importante hétérogénéité de distribution entre les populations humaines pour un grand nombre de variants pharmacogénétiques (Aminkeng et al., 2014; Chen et al., 2010b; Chowbay et al., 2008; Fuselli et al., 2010; Kim et al., 2012; Li et al., 2011; Loh et al., 2013; Man et al., 2010; Ramos et al., 2013; Sistonen et al., 2009; Suarez-Kurtz et al., 2012a; Wilson et al., 2001). Ces variants sont impliqués dans la réponse à de nombreuses molécules utilisées couramment en pratique médicale, telles que la rosuvastatine (Yasuda et al., 2008), les anticoagulants oraux de type antivitamine K (AVK) (Limdi et al., 2010), les bêta-bloquants (Muszkat, 2007), la carbamazépine (Yasuda et al., 2008), les anticancéreux (Aminkeng et al., 2014; Loh et al., 2013; O'Donnell and Dolan, 2009), les antalgiques opiacés, le tacrolimus (Yasuda et al., 2008), etc.

Les études ayant analysé la différenciation des pharmacogènes à une échelle mondiale ont détecté un niveau de différenciation génétique particulièrement élevé pour des gènes d'intérêt majeur dont *VKORC1* (Chen et al., 2010; Kim et al., 2012; Li et al., 2011; Ramos et al., 2013), *ADH1B* (Chen et al., 2010; Kim et al., 2012), et les gènes codant pour les enzymes du cytochrome P450 (Man et al., 2010; Wilson et al., 2001), en particulier *CYP3A4* et *CYP3A5* (Aminkeng et al., 2014; Chowbay et al., 2008; Polimanti et al., 2012; Li et al., 2011; Ramos et al., 2013). Ces gènes sont d'un grand intérêt en pharmacogénétique car ils sont impliqués dans la réponse à de nombreuses molécules importantes utilisées en thérapeutique. Ainsi le gène *VKORC1* code, pour rappel, pour la cible pharmacologique des AVK, médicaments couramment employés à travers le monde ; *ADH1B* participe à la clairance de médicaments importants en clinique tels que les antalgiques opiacés ; et

CYP3A4 et *CYP3A5* à la biotransformation de respectivement plus de la moitié et un tiers des médicaments utilisés en pratique médicale aujourd'hui. *ADH1B* est en outre impliqué dans le métabolisme de l'alcool.

Diversité pharmacogénétique intra-continentale

Par ailleurs, cette hétérogénéité de répartition des variants pharmacogénétiques n'est pas observée uniquement entre continents, mais également entre les populations d'un même continent. Cette observation est particulièrement vraie en Afrique, qui représente le continent présentant la plus grande diversité pharmacogénétique (Aminkeng et al., 2014; Dandara et al., 2014; Ramos et al., 2013; Suarez-Kurtz et al., 2012b). Par exemple Ramos *et al.* (2013) rapportent une valeur moyenne de deltaMAF au sein de huit populations africaines égale à 0,10 pour un ensemble de 1 156 variants de la pharmacocinétique (Ramos et al., 2013). Cette valeur est seulement égale à 0,04 lorsque les trois populations d'Asie de l'Est sont considérées.

C'est effectivement entre les populations d'Asie de l'Est qu'est observée la plus faible différenciation pharmacogénétique (Chen et al., 2010b; Man et al., 2010; Ramos et al., 2013). Cependant il est important de noter que des différences de fréquence sont observées entre les populations du continent asiatique considéré dans son ensemble (Chowbay et al., 2008). C'est par exemple le cas pour l'allèle *UGT1A1*28* qui est associé à un risque accru de toxicité à l'irinotécan (Iyer et al., 2002) et qui, depuis 2005, fait l'objet d'un test pharmacogénétique recommandé par la FDA avant d'initier une thérapie par cet agent anticancéreux (FDA, 2005). L'allèle *UGT1A1*28* est retrouvé à des fréquences bien plus faibles en Asie (~ 10 %) qu'en Europe (~ 30 %) et Afrique (~ 45 %) (Beutler et al., 1998; Hall et al., 1999). Or, à Singapour, alors que cet allèle est retrouvé chez respectivement 19 % et 16 % des individus d'origine malaisienne et chinoise, sa fréquence est près de deux fois plus élevée (35 %) chez des individus d'origine indienne (Balram et al., 2002).

De façon intéressante, il a été observé qu'un variant pharmacogénétique fréquent dans un continent pouvait être retrouvé à une fréquence faible dans une ou plusieurs populations de ce continent, et réciproquement (Ramos et al., 2013).

Ces observations indiquent qu'il n'est pas possible d'extrapoler les fréquences observées d'un continent donné à l'ensemble des populations de ce continent et soulignent la nécessité d'inclure différentes populations d'un même continent dans les études de pharmacogénétique, afin de bien capturer la variabilité génétique des populations humaines pour la réponse aux médicaments.

Diversité pharmacogénétique intra-populationnelle

Au sein d'une population très mélangée comme la population brésilienne, où les individus appartiennent à des origines ethniques diverses (africaine, européenne et amérindienne), des différences importantes de répartition des variants pharmacogénétiques sont observées. Par exemple dans ce pays, la fréquence des allèles situés dans les gènes *CYP2C8* et *CYP2C9* varie en fonction de la région géographique d'origine des individus, de leur appartenance auto-déclarée à un groupe ethnique et des proportions de leur génome d'origine européenne et africaine (Suarez-Kurtz et al., 2012c). En étendant leur analyse à 12 pharmacogènes, les auteurs de cette étude ont mis en évidence une variation spatiale (*i.e.*, entre les différentes régions géographiques du Brésil) du niveau de différenciation génétique entre des individus de différentes origines ethniques (Suarez-Kurtz et al., 2012a). Parmi les variants concernés par cette hétérogénéité géographique et ethnique se trouvent des variants d'importance en pharmacogénétique, situés dans les gènes *ABCB1*, *CYP3A5*, *SLCO1B1*, *SLCO1B13* et *VKORC1* (Suarez-Kurtz et al., 2012b).

En Inde, où la diversité génétique est particulièrement élevée, notamment du fait de différences dans l'histoire démographique des sous-populations, d'une grande diversité socio-culturelle et linguistique et de niveaux élevés d'endogamie, des disparités dans la distribution de variants situés dans les pharmacogènes ont été observées (Indian Genome Variation Consortium, 2008). C'est le cas par exemple des SNPs rs1056827 de *CYP1B1*, rs4147536 de *ADH1B* et rs1042713 de *ADRB2* qui affichent une répartition inégale entre les différents groupes socio-culturels et linguistiques. Ce dernier est associé avec une mauvaise réponse au salbutamol, un antiasthmatique, chez des patients

indiens (Kukreti et al., 2005). En conséquence, le label « indien » classiquement utilisé dans les études génétiques, n'est pas adéquat pour tenir compte des différences observées entre les groupes ethnolinguistiques du pays car il suppose à tort une homogénéité génétique de l'Inde et risque d'entraîner une mauvaise interprétation des résultats des études d'associations réalisées sur la réponse aux médicaments.

Par ailleurs, une diversité pharmacogénétique intra-populationnelle peut être observée y compris dans un pays comme la Finlande, où le degré de mélange des individus de diverses origines ethniques est moindre. C'est le cas des gènes *CYP2C9* et *CYP2C19* dont la distribution entre les régions Est et Ouest du pays est hétérogène (Sistonen et al., 2009), ce qui peut être dû à l'histoire différentielle du peuplement de ces deux régions géographiques : les populations vivant dans la région Est ont été plus affectées par des effets fondateurs et la dérive génétique que celles vivant à l'Ouest du pays (Sajantila et al., 1996).

Cette disparité pharmacogénétique intra-populationnelle soulève des difficultés pour les autorités nationales réglementant l'usage des médicaments, notamment dans la formulation de directives qui soient adaptées à l'ensemble des individus de la population.

Importante diversité pharmacogénétique des populations africaines

De manière générale, les études qui ont exploré le degré de différenciation génétique des populations humaines pour les pharmacogènes ont remarqué que les populations africaines se distinguent le plus des autres populations. Par exemple, les études comparant la différenciation génétique entre paires de populations ont observé un niveau de différenciation plus important (valeurs de F_{ST} plus élevées) lorsqu'une population africaine est incluse dans la comparaison (Chen et al., 2010b; Kim et al., 2012; Li et al., 2011). La différenciation élevée de l'Afrique par rapport au reste du monde est à la fois quantitative (nombre de gènes concernés par une différenciation génétique importante) et qualitative (valeurs de F_{ST} ou de deltaMAF). Cette observation a été faite quel que soit le niveau de regroupement géographique des

individus : continental (Chen et al., 2010b; Kim et al., 2012; Nelson et al., 2012; Yasuda et al., 2008)(Li et al., 2011) ou intra-populationnel (Suarez-Kurtz et al., 2012b) ; mais aussi quel que soit le jeu de pharmacogènes : qu'il s'agisse de un (Fuselli et al., 2010; Wooding et al., 2002) ou de plusieurs gènes (Li et al., 2011; Polimanti et al., 2012) impliqués dans la pharmacocinétique des médicaments ; de un (*cf.* étude de *VKORC1* présentée dans la partie 2 de cette thèse) ou de plusieurs gènes (Nelson et al., 2012) impliqués dans la pharmacodynamie ; ou les pharmacogènes majeurs (gènes VIP) comprenant des gènes de toutes les catégories pharmacogénétiques (Chen et al., 2010b).

Cette grande diversité pharmacogénétique des populations africaines est bien illustrée par l'exemple de l'allèle défectueux *CYP3A5*3*, qui lorsqu'il est porté à l'état homozygote conduit à une absence d'expression de l'enzyme CYP3A5 (Kuehl et al., 2001). Alors qu'il affiche des fréquences homogènes en Europe et en Asie (Dandara et al., 2014), cet allèle est présent en Afrique à des fréquences très variables entre les populations (allant de 4 à 81 %), ainsi qu'au sein de différents groupes ethniques d'une même population (de 29 à 65 % en Éthiopie, de 23 % à 40 % au Cameroun) (Bains et al., 2013). Le CYP3A5 participant à la biotransformation d'un grand nombre de médicaments, notamment des anti-infectieux couramment utilisés en Afrique tels que les antituberculeux, antirétroviraux et antipaludiques, cette hétérogénéité de distribution de l'allèle *CYP3A5*3* est susceptible d'entraîner des différences significative de réponse à ces médicaments dans les populations africaines (Dandara et al., 2014).

De plus, certains variants génétiques susceptibles d'impacter la réponse aux médicaments sont retrouvés exclusivement dans les populations africaines, comme c'est le cas de l'allèle *CYP2D6*17* qui entraîne une activité enzymatique du CYP2D6 diminuée (Masimirembwa and Hasler, 1997). En effet, la diversité génétique du gène *CYP2D6* est supérieure dans les populations africaines que dans les populations européennes (Fuselli et al., 2010). Les relations génotype-phénotype établies dans les populations européennes pour ce gène ne sont pas toujours retrouvées dans les

populations africaines chez qui, de surcroît, on ne cesse de découvrir de nouveaux variants (Dandara et al., 2014). Le polymorphisme du CYP2D6 est notamment susceptible d'influencer la réponse au traitement de nombreux médicaments utilisés en psychiatrie, comme les antipsychotiques et antidépresseurs (Masimirembwa and Hasler, 1997), mais aussi celle du tamoxifène, un anticancéreux utilisé dans le traitement du cancer du sein (Province et al., 2014), et de la primaquine, molécule utilisée dans le traitement du paludisme à *Plasmodium vivax* et *Plasmodium ovale* (Bennett et al., 2013). Un autre exemple concerne la mutation Asp36Tyr du gène *VKORC1*, conférant un phénotype de résistance aux AVK, qui est essentiellement retrouvée dans les populations africaines (Aklillu et al., 2008).

Lien entre variabilité pharmacogénétique et différences inter-populationnelles dans les phénotypes de réponse aux médicaments

L'hétérogénéité de répartition des variants pharmacogénétiques peut expliquer des différences de réponse aux médicaments observées entre populations. Par exemple, une étude se concentrant sur les gènes impliqués dans la réponse aux anticancéreux (anthracyclines, dérivés du platine, fluorouracile, vincristine), antirétroviraux et antimycobactériens (névirapine, rifampicine, éfavirenz, ténofovir) a remarqué que les variants associés à une mauvaise réponse à ces médicaments (efficacité diminuée et/ou toxicité augmentée) étaient plus présents dans les populations africaines que dans les populations européennes, alors que les variants associés à une meilleure réponse étaient plus présents dans ces dernières (Aminkeng et al., 2014). Ces différences génétiques pourraient expliquer le fait que les épisodes de réponse à ces molécules altérée (augmentation du taux de survenue d'effets indésirables, diminution de la réponse thérapeutique) soient plus fréquemment observés dans les populations africaines en comparaison des populations européennes. Par ailleurs, une méta-analyse de la répartition des variants pharmacogénétiques impliqués dans la réponse aux molécules utilisées dans le traitement du cancer colorectal (fluorouracile, irinotécan, oxaliplatine, capécitabine) entre des patients originaires d'Asie de l'Est ou d'Europe a montré que les variants associés à des taux de toxicité plus faibles

étaient plus souvent retrouvés chez les sujets asiatiques. Cette observation peut être reliée au plus faible taux de survenue d'effets indésirables en réponse à ces traitements chez ces sujets (Loh et al., 2013).

Ces observations illustrent l'intérêt de l'étude de la différenciation génétique observée entre les populations humaines pour les pharmacogènes, qui permet d'expliquer une partie de la variabilité inter-populationnelle dans la réponse aux médicaments.

Limites de ces études

Bien qu'ayant permis des découvertes intéressantes sur la compréhension de la variabilité inter-populationnelle de la réponse à certains médicaments, ces études sont également sanctionnées par un certain nombre de limites se rapportant :

- Au nombre et au type de gènes étudiés : en effet certaines études se concentrent sur un seul ou sur un petit nombre de gènes candidats (Fuselli et al., 2010; Sabbagh et al., 2011; Sistonen et al., 2009; Suarez-Kurtz et al., 2012c; Wilson et al., 2001) ; ou se concentrent très majoritairement sur les gènes de la pharmacocinétique (Chowbay et al., 2008; Li et al., 2011; Man et al., 2010; Polimanti et al., 2012; Ramos et al., 2013; Wilson et al., 2001). Ou alors il s'agit de gènes choisis pour leur implication dans la réponse à des molécules bien précises, comme les anticancéreux (Aminkeng et al., 2014; Loh et al., 2013), si bien qu'elles ne permettent pas d'analyser la différenciation génétique des populations humaines pour la réponse aux médicaments en général.
- Aux populations analysées, soit de par leur faible nombre (Kim et al., 2012; Suarez-Kurtz et al., 2012a), soit parce qu'elles sont souvent restreintes à des populations d'origine européenne ou asiatique (Kim et al., 2012; Loh et al., 2013; Man et al., 2010). En effet, comme nous l'avons vu, il est important d'inclure un nombre suffisant de populations dans les études de pharmacogénétique, en particulier dans les régions du monde où une forte hétérogénéité génétique entre populations est observée, comme en Inde ou en Afrique.

- Aux méthodes employées qui ne permettent pas d'identifier correctement les profils de différenciation génétique atypiques. D'une part parce qu'un certain nombre d'études se contentent de fournir une description des fréquences alléliques sans estimer la différenciation génétique des populations avec des indices appropriés tel que le F_{ST} (Aminkeng et al., 2014; Chowbay et al., 2008; Loh et al., 2013; Man et al., 2010; Ramos et al., 2013). D'autre part parce que même lorsque l'indice F_{ST} est employé, les valeurs estimées ne sont pas comparées à des distributions empiriques construites à partir de variants non pharmacogénétiques répartis sur le génome entier, ne rendant pas possible l'estimation de leur significativité statistique (Chen et al., 2010b; Kim et al., 2012; Suarez-Kurtz et al., 2012a). On ne peut ainsi déterminer si les profils de différenciation génétique observés sont extrêmes par rapport à la variation du reste du génome. De plus, les valeurs de F_{ST} estimées ne sont jamais interprétées en fonction de la MAF des variants considérés, qui pourtant, comme nous l'avons expliqué dans le chapitre 2 de la partie 1, influe sur la valeur de F_{ST}.
- Enfin, aux données génétiques utilisées, qui sont en large majorité des données de génotypage ne permettant pas de capturer la totalité de la variation génétique des gènes étudiés, notamment pas les variants rares et peu fréquents. De surcroît, ces données sont affectées par le biais de découverte des variants (*ascertainment bias*) dont nous avons déjà parlé, qui conduit à une représentation biaisée de la variabilité génétique des populations, en particulier à une sous-représentation des variants non présents dans les populations du panel de découverte des variants.

2. Exemples de gènes de la réponse aux médicaments soumis à l'action de la sélection naturelle

L'intérêt d'étudier l'impact de la sélection naturelle sur les gènes de la réponse aux médicaments est double. D'une part cela permet d'améliorer notre compréhension des phénomènes d'adaptation des populations

humaines à leur environnement chimique. D'autre part, étant donné que la sélection naturelle agit sur les phénotypes et cible donc les variants directement impliqués dans la détermination de ces derniers, l'étude des signaux de sélection a le potentiel de déceler des gènes et des variants présentant une différenciation génétique atypique entre les populations, potentiellement responsables d'une partie des différences inter-populationnelles de réponse aux médicaments.

Aujourd'hui, les preuves que la sélection naturelle a agi sur les gènes de la réponse aux médicaments sont nombreuses et concernent aussi bien des gènes de la pharmacocinétique, par exemple certains gènes du cytochrome P450 (Polimanti et al., 2012) dont les gènes de la superfamille *CYP3A* (Thompson et al., 2004) et *CYP1A2* (Wooding et al., 2002) ; que des gènes de la pharmacodynamie, par exemple le gène *DRD2* qui code pour le récepteur de la dopamine (Lao et al., 2007) et le gène *VKORC1* dont nous venons de présenter l'étude détaillée.

L'étude systématique de Li *et al.* (2011) de 283 gènes de la pharmacocinétique a mis en évidence que le partage des signaux de sélection dans les différentes populations d'une même région géographique différait selon la région considérée : alors que les signaux de sélection sont le plus souvent sporadiques en Afrique, ils sont mieux répartis dans les autres régions du monde (Li et al., 2011).

Les pressions de sélection sur les pharmacogènes, nous l'avons vu, pourraient être reliées à des phénomènes d'adaptation des populations en réponse à des changements de l'environnement chimique, notamment alimentaire. Deux études ont ainsi montré que la répartition des variants génétiques des pharmacogènes *CYP2D6* (Fuselli et al., 2010) et *NAT2* (Sabbagh et al., 2011) pouvait être reliée à des différences de mode de vie et d'alimentation.

Bien que les résultats de ces études suggèrent un rôle prépondérant de la sélection naturelle dans la différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments, elles présentent

certaines limites pour différentes raisons relatives aux nombres de gènes explorés, aux populations analysées et aux tests de sélection employés. Par exemple, l'analyse de sélection de Polimanti *et al.* (2012) n'utilise que le test iHS, appliqué dans seulement trois populations sur un faible nombre de variants (Polimanti et al., 2012). L'étude de Li *et al.* (2011) n'emploie que des tests de sélection adaptés à détecter des balayages sélectifs complets ou presque complets sur des données de génotypage (Li et al., 2011). Par ailleurs, rares sont les études qui permettent de démontrer clairement la présence d'une signature génomique de sélection naturelle comme c'est le cas par exemple pour le gène *ADH1B* en Asie de l'Est (Li et al., 2008a) et *NAT2* dans des populations d'Eurasie (Patin et al., 2006a). De telles études requièrent un travail approfondi centré sur le gène, combinant l'analyse de la diversité allélique et haplotypique à une échelle géographique fine à celle de la sélection naturelle qui elle-même nécessite l'emploi de plusieurs tests, si possible complémentaires, appliqués à des données génétiques denses.

Dans l'étude présentée ci-après, nous avons quantifié la différenciation pharmacogénétique de 14 populations mondialement réparties dont trois populations africaines. L'utilisation de données de séquençage nous a permis d'avoir une couverture génétique très dense, favorisant la détection de nouveaux variants pharmacogénétiques d'intérêt qui, de par leur extrême différenciation, pourraient expliquer les différences de réponse aux médicaments entre les populations humaines.
Contrairement à la plupart des études de génétique des populations réalisées sur les gènes de la réponse aux médicaments, nous avons utilisé une approche exhaustive analysant l'ensemble des gènes majeurs impliqués dans la réponse aux médicaments, sans nous concentrer sur des gènes impliqués dans la réponse à des médicaments particuliers, ou appartenant à une catégorie pharmacogénétique particulière. Notre étude nous permet ainsi de couvrir des gènes associés à la réponse à un grand nombre de molécules et impliqués dans les différentes phases de la réponse aux médicaments.

.

Chapitre 2

Analyse de la différenciation génétique des populations humaines pour les pharmacogènes majeurs et rôle de la sélection positive

Nous présentons dans ce chapitre l'analyse des profils de différenciation génétique entre 14 populations humaines pour les gènes majeurs de la réponse aux médicaments. Nous nous sommes intéressés dans ce travail aux profils de différenciation extrêmes (c'est-à-dire aux valeurs de F_{ST} les plus élevées), que nous avons cherché à expliquer par la détection de la sélection positive.

1. Résumé de l'article 2

L'étude que nous présentons dans les pages qui suivent a pour objectifs :

(1) de décrire les profils de différenciation génétique des populations humaines pour les gènes de la réponse aux médicaments ;

(2) d'évaluer comment ces profils diffèrent en fonction de la catégorie pharmacogénétique à laquelle appartiennent ces gènes ;

(3) d'identifier les variants génétiques qui présentent un profil de différenciation génétique atypique par rapport à la variation globale du génome ;

(4) d'évaluer le rôle de la sélection dans la détermination de ces profils de différenciation extrêmes,

(5) de fournir une liste de nouveaux variants pharmacogénétiques, qui pourraient expliquer une partie des différences de réponse aux médicaments observés entre les populations et les individus.

Pour conduire cette étude, nous avons choisi les 45 gènes majeurs impliqués dans la réponse aux médicaments : les gènes VIP, sélectionnés par les investigateurs du PGRN au travers de la base de données PharmGKB (http://www.pharmgkb.org/)[9].

Nous avons, de façon systématique, quantifié le niveau de différenciation génétique global de l'ensemble des variants de type SNV des 45 gènes VIP identifiés dans le Projet 1000 Génomes, en utilisant l'indice de fixation F_{ST}. Celui-ci a été calculé à deux niveaux : entre continents (F_{ST} intercontinental) et entre populations (F_{ST} inter-populationnel). Afin de déterminer si les valeurs de F_{ST} obtenues pour ces variants reflétaient une différenciation génétique atypique par rapport à la variation de fond du génome, nous les avons comparées à celles d'une distribution génome entier incluant 25 millions de variants génétiques indépendants de la base de données 1000 Génomes, en tenant compte de leur position dans le gène et de leur effet sur la protéine. Nous avons ainsi calculé deux *P*-values empiriques pour chacun des variants de la réponse aux médicaments : une correspondant à la distribution du F_{ST} génome entier (distribution empirique principale), et une correspondant à la sous-distribution n'incluant que les variants ayant la même annotation fonctionnelle que le variant testé (sous-distribution empirique). Nous avons considéré qu'un variant présentait un profil de différenciation génétique extrême si ces deux *P*-values étaient inférieures au seuil de signification 0,05.

Au total, nous avons identifié 636 variants présentant un profil de différenciation géographique extrême parmi les 9 695 variants étudiés, soit une proportion de 6,6 %, ce qui représente un excès par rapport à la proportion observée pour l'ensemble des variants du génome (3,1 %, $P < 10^{-8}$). Parmi eux figurent 12 variants connus pour jouer un rôle important dans la

[9] Au moment où nous avons constitué cette liste en juin 2012, elle contenait en réalité 46 gènes VIP. Nous avons exclu de nos analyses le gène *G6PD*, situé sur le chromosome X. Cette liste régulièrement actualisée comprend aujourd'hui 51 gènes, incluant les nouveaux gènes *CYP2C8*, *CYP2E1*, *EGFR*, *NAT2* et *SLC22A1*. Le gène *NAT2* fera l'objet d'une étude à part entière, présentée dans la partie 4 de ce document.

réponse aux médicaments, annotés comme « variants clés » dans la base PharmGKB.

Afin d'évaluer si le degré de différenciation génétique différait selon le rôle des gènes VIP dans la réponse aux médicaments, nous avons comparé la proportion de variants présentant un profil de différenciation génétique atypique entre les différentes catégories pharmacogénétiques à l'aide de tests non paramétriques. De manière intéressante, nous avons observé que cette proportion était supérieure pour les variants de la pharmacodynamie par rapport à ceux de la pharmacocinétique (7,4 % *vs* 6,2 %, $P = 0,027$). Des différences ont également été observées entre les différentes catégories pharmacocinétiques, avec un excès de variants extrêmement différenciés dans les catégories « métabolisme de phase II » et « gènes modificateurs » (respectivement 10,2 % et 15,9 %), en comparaison avec les catégories « métabolisme de phase I » et « transport » (respectivement 5,3 % et 6,2 %).

L'application d'un outil bioinformatique de prédiction de l'effet fonctionnel des 636 variants présentant un profil de différenciation géographique atypique nous a permis d'identifier 35 variants ayant une forte probabilité d'avoir un impact sur la fonction ou l'expression de la protéine produite. Ces variants constituent donc de bons candidats pour intervenir dans la variabilité de réponse aux médicaments observée entre les populations et les individus.

Ces 35 variants, ainsi que les 12 variants clés précédemment identifiés, appartiennent à 13 gènes VIP. Pour chacun d'eux, nous avons recherché la présence de signatures moléculaires de la sélection positive, en appliquant différentes méthodes spécialement adaptées à la détection de ce type de sélection : les tests iHS, XP-EHH et XP-CLR. Nous avons détecté un signal de sélection pour 11 des 13 gènes étudiés, révélant le rôle capital de la sélection naturelle dans les profils de différenciation géographique des variants pharmacogénétiques et, probablement, dans la détermination de la variabilité de réponse aux médicaments observée aujourd'hui entre les populations humaines.

2. Article 2

Patillon B, Luisi P, Letort S, Laayouni H, Génin E, Sabbagh A. *Global patterns of population genetic differentiation for genes involved in drug response*. En cours de rédaction.

Les parties Matériel & Méthodes et Résultats sont présentées ci-après dans un format article et sont rédigées en anglais. La partie Discussion, rédigée en français, fait l'objet de la dernière partie (2.3) de cette partie 3. Les découvertes et limites de notre étude y sont discutées, et les résultats concernant les 13 gènes d'intérêt sont synthétisés à travers un tableau bilan. Les Figures et Tables supplémentaires sont présentées en Annexe 2.

2.1. Matériel & Méthodes

Data retrieval

Whole-genome variation data generated by the 1KG project in 1,089 unrelated individuals was directly downloaded from the 1000 Genomes ftp site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/), using the phase 1 integrated release version 3 of April 2012 (The 1000 Genomes Project Consortium et al., 2012). The 1,089 individuals are drawn from 14 different populations in sub-Saharan Africa, Europe, East Asia, and the Americas (Figure S1): Yoruba in Ibadan, Nigeria (YRI); Luhya in Webuye, Kenya (LWK); people with African ancestry in the Southwest United States (ASW); Utah residents with Northern and Western European ancestry (CEU); Tuscans in Italy (TSI); British in England and Scotland (GBR); Finnish in Finland (FIN); Iberians in Spain (IBS); Han Chinese in Beijing, China (CHB); Southern Han Chinese in China (CHS); Japanese in Tokyo, Japan (JPT); people with Mexican ancestry in Los Angeles, California (MXL); Colombians in Medellin, Colombia (CLM); and Puerto Ricans in Puerto Rico (PUR). From the vcf (variant call format) files, we extracted exclusively the low-coverage SNV calls obtained using VQSR (Variant Quality Score Recalibrator method) in order to avoid any bias that might result from differences between low-coverage whole-genome calls and high-coverage exome SNV calls. A total of 36,382,866 autosomal SNVs were considered for analysis after exclusion of indels and variants that were not seen at least twice in the 1,089 individuals (*i.e.,* singletons). A functional annotation of SNVs was performed using classification from the dbSNP database (build 137) (http://genome.ucsc.edu/cgi-bin/hgTables:SNV137.txt). SNVs were assigned to two main classes: genic and nongenic SNVs. Genic SNVs were further classified as intronic, 5'-UTR, 3'-UTR, coding synonymous, coding non-synonymous or splice-site. Derived allele frequencies (DAF) were calculated using ancestral allele information provided by the 1000 Genomes Consortium (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/).

Selection of genes and variants involved in drug response

Pharmacogenes were chosen using the Pharmacogenomics Knowledge Base (PharmGKB; www.pharmgkb.org) (Hewett et al., 2002; Klein et al., 2001) which provides a list of the most important genes involved in drug response. In June 2012, a list of 46 genes, annotated as Very Important Pharmacogenes (VIPs) by the investigators of the Pharmacogenetic Research Network (PGRN), was available. We discarded *G6PD* located on the X chromosome. Analyses were performed on the whole set of genes (N=45) as well as subsets defined according to the two pharmacogenetic categories: pharmacokinetics (PK) (N=25) and

pharmacodynamics (PD) (N=20). PK genes were further subdivided into phase I metabolism (N=14), phase II metabolism (N=5), transporter (N=3) and modifier (N=3) genes (Figure 3.1 and Table S1). Within each VIP gene, key variants of known clinical relevance are listed in PharmGKB through the VIP summary (Hernandez-Boussard et al., 2008, Mc Donagh et al., 2011). These variants were selected by PGRN members and are annotated with evidence of either *in vitro* or *in vivo* functional effect on drug response. Albeit not exhaustive, this list contains the most commonly studied variants and those repeatedly associated with drug effect (Peters and McLeod, 2008). We considered all 1KG variants occurring within each VIP gene and their 2-kb flanking regions, resulting in a total of 9695 SNVs including 90 key variants (Table S1, Figures S2 and S3).

Population genetic differentiation

Global levels of population genetic differentiation in the 45 VIP genes were evaluated by using the fixation index F_{ST} (Wright, 1951), which quantifies the proportion of genetic variance explained by allele frequency differences among populations. F_{ST} ranges from 0 (for genetically identical populations) to 1 (for completely differentiated populations). F_{ST} scores were computed for the 9695 VIP variants using the BioPerl module PopGen (Stajich et al. 2002) at two different levels: (i) among populations (inter-population F_{ST}) and (ii) among continental regions (inter-continental F_{ST}), *i.e.* when the individual samples were grouped into major geographical regions (Africa, Europe, East Asia, and America) according to their predominant component of ancestry.

Extreme values of F_{ST} can result from natural selection but also from nonselective events such as demographic changes and genetic drift. Because such nonselective processes randomly act on the genome, they are expected to have the same average effect across the genome, in contrast to natural selection, which impacts population differentiation in a locus-specific manner. The genome-wide variation data provided by the 1KG project can thus be used to infer the action of natural selection by adopting an outlier approach (Kelley et al. 2006). For that purpose, we built nine empirical distributions of the F_{ST} statistic by considering different subsets of SNVs defined according to their physical location and/or functional impact. To obtain distributions of likely independent observations, a LD-based pruning procedure was applied to each of these nine subsets using Plink (Purcell et al., 2007) with default parameters (pruning based on a variance inflation factor of at least 2 within each sliding window of 50 SNVs with a step of five SNVs). This resulted for instance in a total of 25,532,386 independent autosomal SNVs included in the genome-wide empirical distribution. Summary statistics of the nine empirical distributions (genome-wide, nongenic, genic, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice-site) are

provided in Table S2. These distributions were then used as references to assess whether the patterns of genetic differentiation observed at the 9695 VIP variants were atypical. Since we were interested in detecting signals of positive selection, we focused on the upper tail of the F_{ST} distributions. Therefore, the estimated *P*-value corresponds to the proportion of F_{ST} scores in the empirical distribution that are higher or equal to the value observed at the locus of interest. Since F_{ST} strongly correlates with heterozygosity (Barreiro et al., 2008; Beaumont and Nichols, 1996; Elhaik 2012), empirical *P*-values were calculated within bins of SNVs grouped according to their global MAF. A total of 27 bins were considered for the whole MAF range: 10 bins of size 0.001 for MAF between 0 and 0.01, 9 bins of size 0.01 for MAF between 0.01 and 0.10, and 8 bins of size 0.05 for MAF between 0.10 and 0.50.

Linkage disequilibrium analyses

Strength of LD between pairs of markers was measured as r^2 (Hill, 1968) using Plink software (Purcell et al., 2007).

In-silico prediction of SNV's functional effects

The F-SNP method (http://compbio.cs.queensu.ca/F-SNP/) was applied to assess the potential functional effect of SNVs (Lee and Shatkay, 2008). This integrative scoring method combines assessments from 16 independent computational tools and databases, using a probabilistic framework that takes into account both the certainty of each prediction and the reliability of the different tools depending on the physical and functional annotation of the specific variant tested. It provides a functional significance (FS) score that quantitatively measures the possible deleterious effect of the tested SNV at the splicing, transcriptional, translational and post-translational levels. An FS score of 0.5 is considered as the cutoff point for predicting a deleterious effect (Lee and Shatkay, 2009).

Detection of signatures of positive selection

To search for genetic footprints of positive selection in VIP genes, we used three complementary approaches based on the site frequency spectrum (Tajima's *D*), allele frequency differentiation (XP-CLR) and local haplotype structure (iHS and XP-EHH). Tajima's *D* (Tajima 1989) is a classical neutrality test that compares estimates of the number of segregating sites and the average number of pairwise differences between nucleotide sequences (π). A zero value of the test statistic *D* is expected under the null hypothesis of selective neutrality, whereas a positive value (excess of low-frequency variants) is considered as indicative of positive selection. The XP-CLR test (Chen et al. 2010) identifies selective sweeps in a population by detecting significant allele frequency differentiation in an extended

genomic region surrounding the locus of interest as compared to a reference population. Finally,two methods based on the extended haplotype homozygosity (EHH) measure, *i.e.* the sharing of identical alleles across relatively long distances by most haplotypes in a population sample (Sabeti et al. 2002), were applied. The first method, the integrated haplotype score (iHS) (Voight et al. 2006) compares the rate of EHH decay observed for both the derived and ancestral alleles at each core SNV. An extremely positive or negative value at the core SNV provides evidence of positive selection with unusually long haplotypes carrying the ancestral or the derived allele, respectively. The second method, the XP-EHH statistic (Sabeti et al. 2007) compares the integrated haplotype score computed in a test population versus that of a reference population. All selection tests were performed in three different populations of West African, Northern European and East Asian ancestry (YRI, CEU, CHB) using the 1000 Genomes Selection Browser 1.0 (http://hsb.upf.edu) (Pybus et al. 2014). Briefly, Tajima's *D* was calculated using a sliding window approach with a window size of 30 kb and a 3 kb offset. XP-CLR scores were computed at regularly spaced grid points (every 4 kb) using SNV genotypes within overlapping windows of 0.2 cM around each grid point (maximal number of SNVs within each window set to 300). iHS and XP-EHH scores were calculated at each polymorphic position. Statistical significance of the different test statistics was assessed using an empirical approach based on the genome-wide distribution of the computed scores (Pybus et al. 2014). We considered the presence of a positive selection signal in a given gene if (a) at least 10% of the computed scores were significant at the 0.05 genome-wide significance threshold for at least one of the four selection tests (with a minimum number of significant scores set to four) or if (b) at least 5% of the computed scores (with a minimum of four scores) were significant at the 0.05 genome-wide significance threshold for at least two of the four selection tests.

2.2. Résultats et Figures

Detection of unusual patterns of genetic differentiation

We investigated the global patterns of genetic differentiation across 14 worldwide populations in four continents for 9695 variants of 45 major genes involved in human drug response, using the F_{ST} index computed at the inter-population and inter-continental levels. To determine whether these patterns are atypical compared to the remaining variation of the genome, *P*-values were estimated from empirical distributions built from the genome background using the whole-genome variation data provided by the 1KG project (The 1000 Genomes Project Consortium et al., 2012) (see Materials and Methods). Each genetic variant was assigned a 'main *P*-value' derived from the genome-wide empirical distribution and a 'subset *P*-value' derived from the distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, nongenic, near gene, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice site). We considered that any variant having both the main and subset *P*-values below 0.05 displays an unusual pattern of genetic differentiation. Full results for the 9695 VIP variants are provided in Table S3.

Considering the inter-population F_{ST}, an atypical pattern of geographic differentiation was observed for 636 variants out of the 9695 (6.6%) (Figure 3.2). Very similar results were obtained with the inter-continental F_{ST} (Table S3), with a high correlation between the two F_{ST} indices (Pearson's correlation coefficient = 0.992, *P*-value < 10^{-5}) (Figure S4A). It is interesting to note that most variants (79%) with a significant inter-continental F_{ST} score have also a significant inter-population F_{ST} (Figure S4B). We therefore decided to focus on the inter-population F_{ST} results in the following analyses.

The proportion of 6.6% of variants with a significant inter-population F_{ST} (with both the main and subset *P*-values < 0.05) is well above the proportion observed at the genome-wide level (3.1%), denoting an excess of highly differentiated variants in genes involved in drug response (chi-square test with 1 degree of freedom (df = 410, *P*-value = 3 x 10^{-91}) (Figure 3.3A). This is particularly true for variants located in PD genes (7.4%) as compared to PK genes (6.2%), with a significant difference between these two proportions (chi-square test with 1 df = 4.9, *P*-value = 0.028 (Figure 3.3A). The F_{ST} and empirical *P*-values distributions were also significantly different between PD and PK variants (Kolmogorov-Smirnov test, *P*-values = 1.02 x 10^{-5} and 0.015 for the inter-population F_{ST} and main empirical *P*-value, respectively). A significant heterogeneity in the proportion of highly differentiated variants was also observed among the four PK subcategories (phase I metabolism, phase II metabolism, transporter and modifier genes) (chi-square test with 3 df = 63.9, *P*-value = 8.7 x

10^{-14}). While all four categories exhibited proportions significantly higher than the genome-wide average (all *P*-values < 10^{-8}), genes involved in phase II metabolism and modifier genes showed remarkably high values (10.2% and 15.9%, respectively) compared to genes involved in phase I metabolism and transport (5.3% and 6.2%, respectively) (Figures 3.3A and 3.3B). Similar results were obtained when considering the distributions of F_{ST} and empirical *P*-values, with a significant heterogeneity among the four PK subcategories (Kruskal-Wallis test, *P*-value < 2.2 x 10^{-16} for both F_{ST} and main *P*-value distributions).

Interestingly, 12 out of the 90 VIP key variants (13.3%) are present among the 636 highly differentiated SNVs (Figure 3.2). These 12 key variants are located in eight VIP genes involved in both PK (*ADH1A*, *ADH1B*, *CYP2D6*, *CYP3A4* and *P2RY1*) and PD processes (*DRD2*, *VDR* and *VKORC1*) (Table 3.1 and Figure 3.4). Given the known functional impact of these polymorphisms, their peculiar geographical distribution is likely to explain some of the observed population heterogeneity in drug treatment effects. They are all common in humans (global MAF ≥ 0.20), except the low-frequency (global MAF ≤ 0.05) nonsynonymous substitutions in *CYP2D6* (rs61736512, rs28371706 and rs59421388) and *ADH1B* (rs2066702) (Table 3.1). The range of variation of the DAF across the 14 worldwide populations (ΔDAF) greatly varies among variants, ranging from 0.16-0.23 for the three *CYP2D6* SNVs to 0.93 for the two *VKORC1* SNVs (Figure 3.4). Note the high allelic correlation between the two key variants rs61736512 and rs59421388 in *CYP2D6* (r^2 = 0.97 in the global sample of 1089 individuals) and between the promoter rs9923231 and intronic rs9934438 variants in *VKORC1*, in complete LD with each other (r^2 = 1.0). The highest inter-population F_{ST} values are observed for the promoter variant rs2740574 in *CYP3A4* (F_{ST} = 0.60, subset *P*-value = 0.0006) and for the nonsynonymous substitution rs1229984 in *ADH1B* (F_{ST} = 0.59, subset *P*-value = 0.002). These high F_{ST} scores are explained by a significantly lower frequency of the derived allele T of rs2740574 in African populations as compared to non-Africans (0.24 *vs* 0.96, chi-square test 1df = 1230, *P*-value = 1.7 x 10^{-269}) and by a significantly higher frequency of the derived allele T of rs1229984 in East Asians as compared to other world populations (0.73 *vs* 0.03, chi-square test 1 df = 1240, *P*-value = 1.2 x 10^{-271}). They are then followed by the two linked variants rs9923231 and rs9934438 in *VKORC1* (F_{ST} = 0.39, subset *P*-value = 0.004), which exhibit a wide variation in the DAF between East Asians and Africans (0.92 *vs* 0.07, respectively, chi-square test 1df = 766, *P*-value = 1.5 x 10^{-168}). For seven of the 12 key variants, Africa appears as the most differentiated continental region, followed by East Asia for four other variants. Interestingly, American populations exhibit allele frequencies very similar to those observed in Europeans, probably as a result of the high level of European admixture in these samples (1000 Genomes Project Consortium et al., 2012). The highest intra-continental variation in allele frequencies

is observed in Africa. The Iberian Spanish population differs markedly from other Europeans for two key variants (rs975833 in *ADH1A* and rs6277 in *DRD2*).

Identification of putative deleterious variants

To distinguish among the 636 variants showing an unusually strong pattern of geographic differentiation, those that are the most likely to affect protein structure or function, we used the F-SNP tool (Lee and Shatkay, 2008, 2009) to evaluate putative functional effects with respect to four major bio-molecular functions (splicing, transcription, translation and post-translational modification). A total of 35 variants in 11 VIP genes (*ADH1A*, *ADH1B*, *AHR*, *CYP2B6*, *CYP2C9*, *CYP2D6*, *DRD2*, *F5*, *GSTT1*, *P2RY1*, and *VKORC1*) exhibited an FS score $\geq$ 0.5, suggesting a potentially deleterious effect (Table 3.2 and Figure S5). These variants are good candidates to explain between-population variation in drug response phenotypes. They include four of the 12 key variants mentioned before (Table 3.2). Most of them (26 out of 35) have a global MAF $\geq$ 0.05. Two variants in *ADH1B*, located in the upstream regulatory region of the gene, display the highest inter-population F_{ST} values: rs3811801 (F_{ST} = 0.56, subset *P*-value = 0.001) and rs2070898 (F_{ST} = 0.40, subset *P*-value = 0.007). The derived alleles of both SNPs are significantly more frequent in East Asia (Figure 3.5), while not being in significant LD with each other. In particular, the rs3811801 A allele occurs at a high frequency in East Asia (0.61) whereas it is totally absent in other parts of the world. Interestingly, this SNV is in high LD (r^2 = 0.70) with the *ADH1B* rs1229984 key variant previously identified as a highly differentiated variant (F_{ST} = 0.59, subset *P*-value = 0.002, Table 3.1) but not selected here because of its low FS score (FS score = 0.301). Several other variants predicted as deleterious by F-SNP are in significant and strong LD with at least one key variant located in the same gene (Table 3.2), suggesting that the unusual geographic differentiation patterns observed at these variants may not be independent of that/those identified at the key variant(s). This is the case for the intronic rs3819197 SNV of *ADH1A*, predicted to affect the transcriptional regulation of the gene (FS score = 0.50): this variant is in complete LD (r^2 = 1.0) with the intronic rs975833 key variant of *ADH1A*, showing a similarly high inter-population F_{ST} value (F_{ST} = 0.30, subset *P*-value = 0.03) but not selected by F-SNP (FS score = 0.242). Likewise, the two linked *CYP2D6* variants rs1080984 and rs1080986 (r^2 = 1.0), are in complete or near complete LD with two key variants of this gene (r^2 = 0.97 with rs59421388 and r^2 = 1.0 with rs61736512). Finally, seven of the eight variants identified in *DRD2* showed a strong allelic correlation (r^2 values ranging from 0.72 to 0.97) with the coding synonymous rs6277 key variant. Most of them exhibited a higher and more significant F_{ST} value than rs6277, along with a higher FS score (0.50 *vs* 0.23). Apart from these 11 SNVs in *ADH1A*, *ADH1B*, *CYP2D6* and *DRD2*, all other variants presented an

unusually high level of genetic differentiation independently of the known key variant(s) located in the same gene, suggesting they may be important new candidates for functional variation in VIP genes. For some of them, the inter-population F_{ST} score is rather low but unusually high when considering the global MAF of the variant. This is the case for instance for the *CYP2B6* 3'UTR variant (rs28399501; F_{ST} = 0.08, subset *P*-value = 0.003), whose derived allele is exclusively found in the Finnish population where it reaches a frequency of 0.07. As noted previously, differences in allele frequencies with one or several African populations are the most often involved in the high population genetic differentiation observed (21 out of 35 variants).

Detection of signatures of positive selection

To determine to what extent natural selection has influenced the high levels of genetic differentiation observed in the 13 VIP genes identified above (listed in Tables 3.1 and 3.2), we applied four selection tests based on the site frequency spectrum, allele frequency differentiation and local haplotype structure. For each gene, we looked for signatures of positive selection in the main continental regions (Africa, Europe and Asia represented by the YRI, CEU and CHB populations, respectively) where the derived allele of the highly differentiated SNV(s) was more frequent. A signal of positive selection was identified in 11 of the 13 VIP genes in at least one population sample (Table 3.3 and Figure S6). A total of eight signals of selection were detected in the YRI sample (*ADH1A*, *ADH1B*, *AHR*, *CYP2B6*, *CYP2C9*, *DRD2*, *F5* and *VKORC1*), six in CHB (*ADH1A*, *ADH1B*, *CYP3A4*, *F5*, *P2RY1* and *VKORC1*), and four in CEU (*CYP3A4*, *DRD2*, *GSTT1* and *P2RY1*). These findings support an important role of local adaptive events in shaping the present-day distribution of known or potential functional polymorphisms in the main genes involved in drug response.

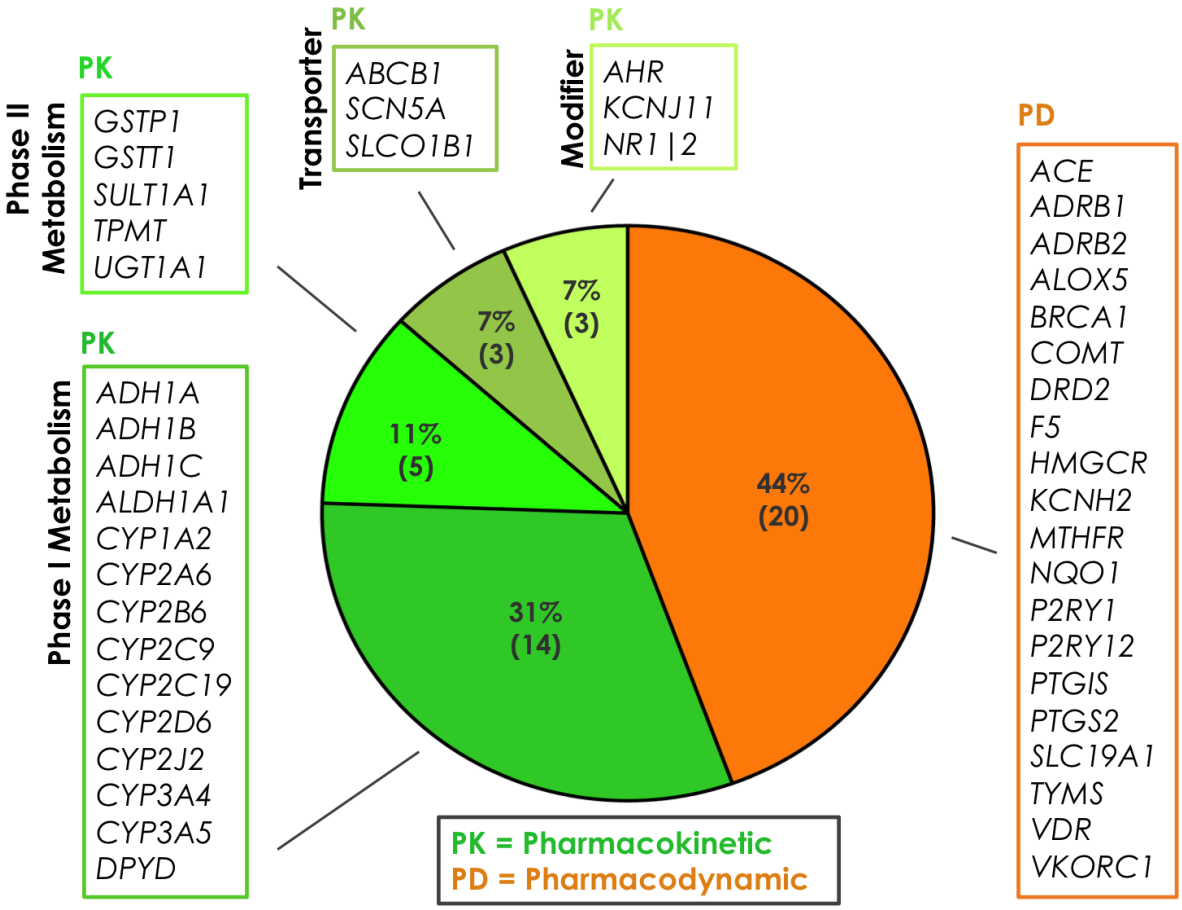


Figure 3.1 | The 45 Very Important Pharmacogenes (VIP) selected. The number of genes in each pharmacogenetic category is indicated in parentheses.

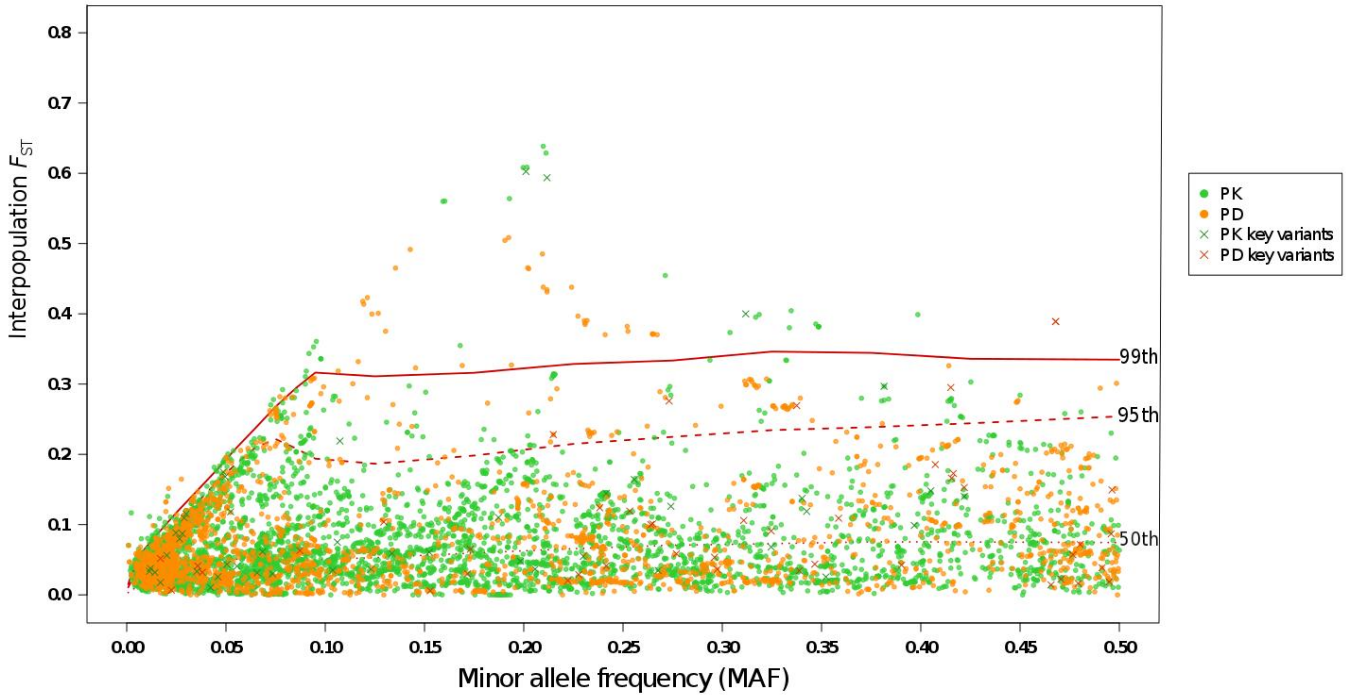
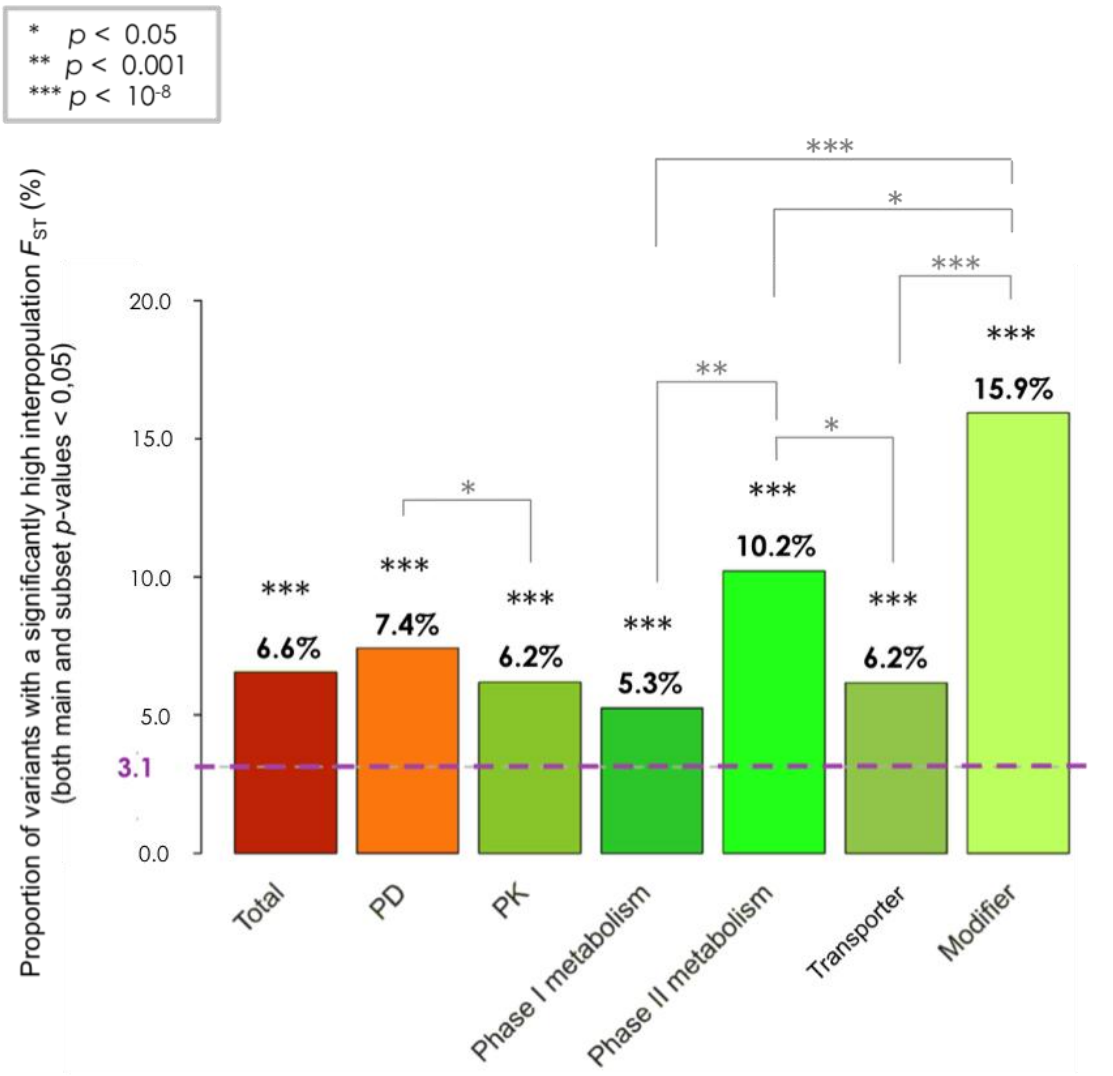


Figure 3.2 | Patterns of population genetic differentiation for the 9695 genetic variants in the 45 VIP genes. The genome-wide empirical distribution of inter-population F_{ST}, computed across the 14 worldwide populations from the 1000 Genomes project, was constructed from 25,532,386 independent autosomal SNVs. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively. Individual F_{ST} values for the 9695 genetic variants are plotted against their global minor allele frequency (MAF). Variants in pharmacokinetic (PK) and pharmacodynamic (PD) genes are shown in green and orange, respectively and key variants are indicated by crosses.

A



B

Chi-square test	Phase I metabolism	Phase II metabolism	Transporter	Modifier
Phase I metabolism	-	-	-	-
Phase II metabolism	2.97×10^{-4}	-	-	-
Transporter	0.181	0.011	-	-
Modifier	1.44×10^{-13}	0.046	1.00×10^{-8}	-

Figure 3.3 | (A) Proportion of variants with a significantly high inter-population F_{ST} in the whole set of 45 VIP genes and in the different pharmacogenetic categories. The dashed purple line indicates the proportion of highly differentiated variants observed at the genome-wide level (3.1%). All pharmacogenetic categories display a significant excess of highly differentiated variants as compared to the genome-wide average (chi-square test x df = XX, all *P*-values < 10^{-8}). Significant pairwise comparisons between the different pharmacogenetic categories are shown in gray (chi-square test of homogeneity with 1 df). **(B) Pairwise comparisons between the different pharmacokinetic subcategories.** *P*-values of the 1df chi-square test of homogeneity comparing the number of variants with significant and non significant F_{ST} are reported.

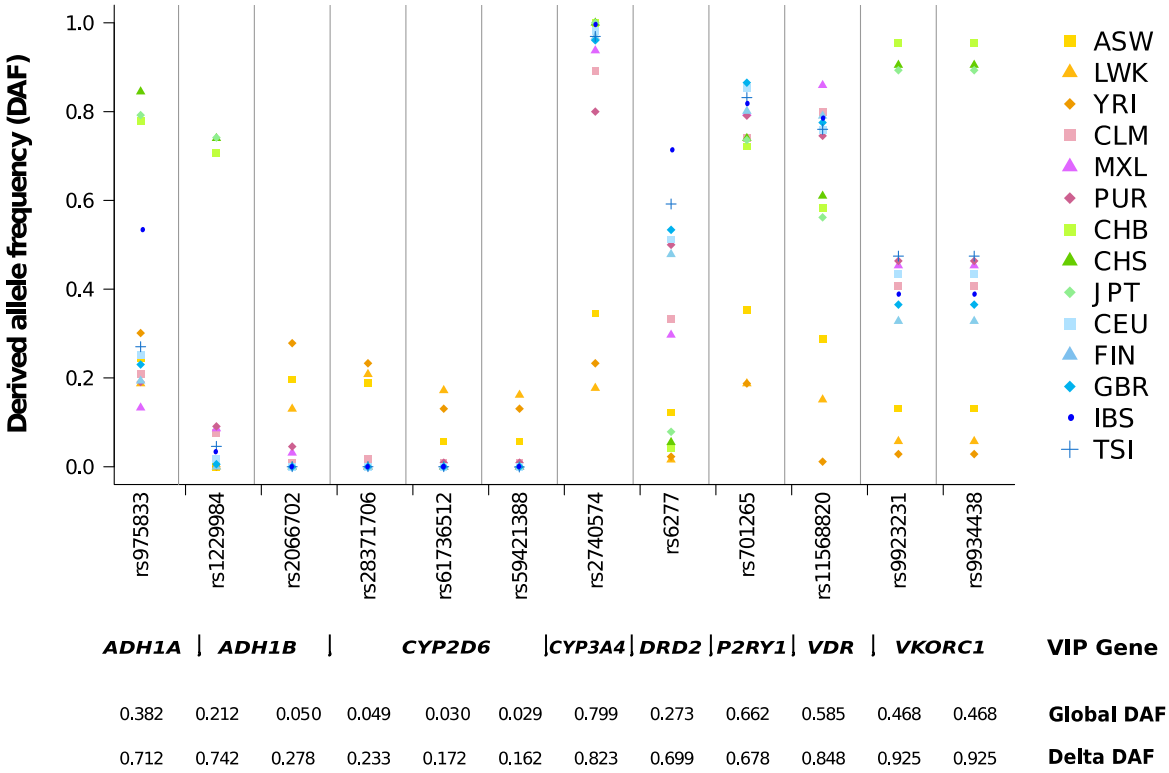


Figure 3.4 | Derived allele frequency (DAF) distribution of the 12 VIP key variants with a significantly high inter-population F_{ST} across the fourteen 1000 Genomes populations. The gene name, global DAF and delta DAF (difference between the highest and the lowest DAF across the 14 worldwide populations) are indicated below each variant.

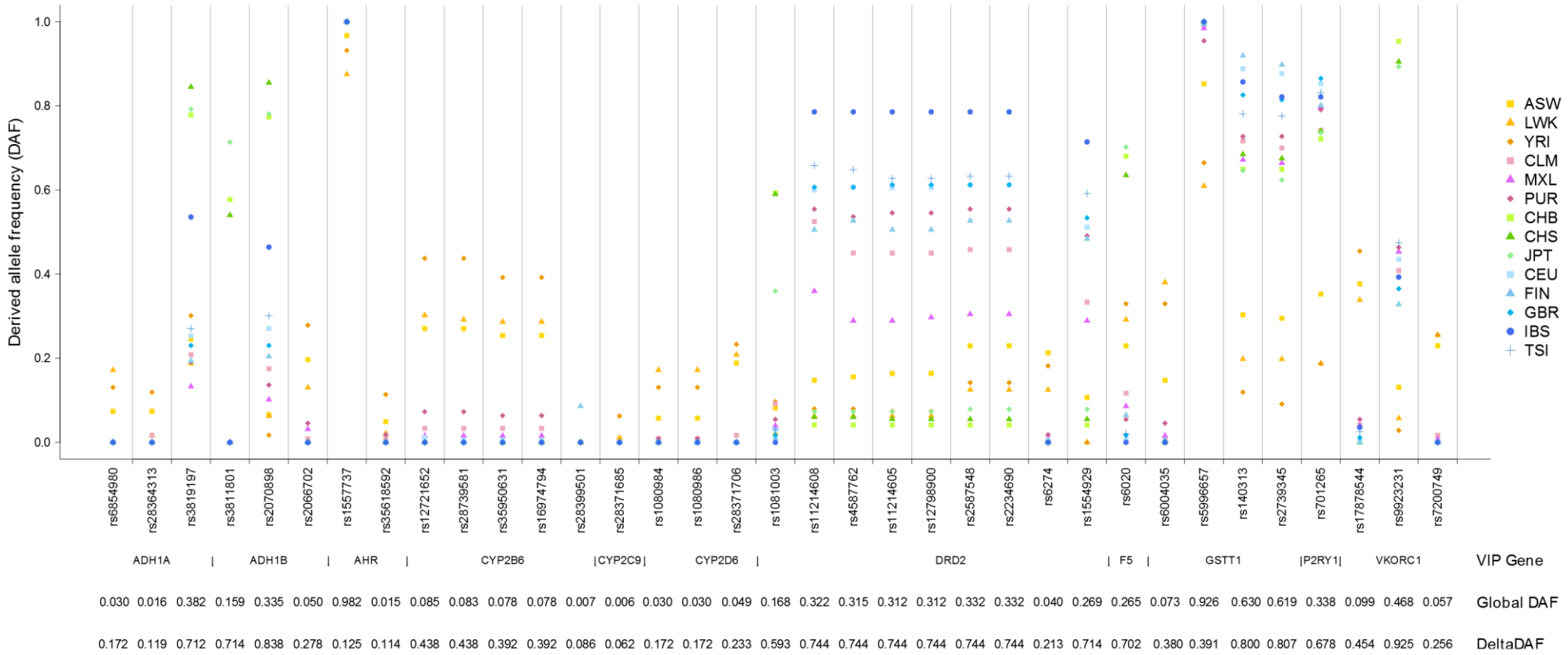


Figure 3.5 | Derived allele frequency (DAF) distribution of the 35 VIP variants with a significantly high inter-population F_{ST} value and a functional significance score FS above 0.5. The gene location, global DAF and delta DAF(difference between the highest and the lowest DAF across the 14 worldwide populations) are indicated below each variant.

Table 3.1 | The 12 VIP key variants with a significantly high interpopulation F_{ST} value.

Key variant	Gene	Pharmacogenetic category	Functional annotation	Global MAF	Interpopulation F_{ST}	F_{ST} main *P*-value[a]	F_{ST} subset *P*-value[b]
rs975833	*ADH1A*	Phase I metabolism	Intronic	0.38	0.30	0.020	0.030
rs1229984	*ADH1B*	Phase I metabolism	Coding nonsynonymous	0.21	0.59	0.0001	0.002
rs2066702	*ADH1B*	Phase I metabolism	Coding nonsynonymous	0.05	0.17	0.024	0.037
rs61736512	*CYP2D6*	Phase I metabolism	Coding nonsynonymous	0.03	0.12	0.012	0.011
rs28371706	*CYP2D6*	Phase I metabolism	Coding nonsynonymous	0.05	0.17	0.013	0.021
rs59421388	*CYP2D6*	Phase I metabolism	Coding nonsynonymous	0.03	0.11	0.019	0.020
rs2740574	*CYP3A4*	Phase I metabolism	Near gene	0.20	0.60	0.0001	0.0006
rs6277	*DRD2*	Pharmacodynamic	Coding synonymous	0.27	0.28	0.023	0.045
rs701265	*P2RY1*	Phase I metabolism	Coding synonymous	0.34	0.27	0.030	0.045
rs11568820	*VDR*	Pharmacodynamic	Intergenic	0.42	0.30	0.021	0.021
rs9923231	*VKORC1*	Pharmacodynamic	Near gene	0.47	0.39	0.003	0.004
rs9934438	*VKORC1*	Pharmacodynamic	Intronic	0.47	0.39	0.003	0.004

Abbreviation: MAF, minor allele frequency.

[a] *P*-values are derived from the genome-wide empirical distribution of interpopulation F_{ST}.

[b] *P*-values are derived from the empirical distribution of interpopulation F_{ST} computed for SNVs with the same functional annotation than the SNV of interest.

Table 3.2 | List of VIP variants with a significantly high inter-population F_{ST} value and a functional significance (FS) score above 0.5.

Gene	SNV	Pharmacogenetic category	Functional annotation	Global MAF	Interpopulation F_{ST}	F_{ST} main P-value[a]	F_{ST} subset P-value[b]	FS score	Key variant
ADH1A	rs6854980	Phase I metabolism	Intronic	0.03	0.12	0.008	0.008	0.50	no
	rs28364313	Phase I metabolism	Intronic	0.02	0.08	0.017	0.017	0.50	no
	rs3819197	Phase I metabolism	Intronic	0.38	0.30	0.020	0.030	0.50	no
ADH1B	rs3811801	Phase I metabolism	Near gene	0.16	0.56	0.0003	0.001	0.50	no
	rs2070898	Phase I metabolism	Near gene	0.33	0.40	0.004	0.007	0.50	no
	rs2066702	Phase I metabolism	Coding nonsynonymous	0.05	0.17	0.024	0.037	0.86	yes
AHR	rs1557737	Modifier	3' UTR	0.02	0.08	0.012	0.012	0.50	no
	rs35618592	Modifier	3' UTR	0.02	0.07	0.048	0.050	0.50	no
CYP2B6	rs28399501	Phase I metabolism	3' UTR	0.01	0.08	0.002	0.003	0.50	no
	rs12721652	Phase I metabolism	Near gene	0.08	0.28	0.017	0.035	0.50	no
	rs28739581	Phase I metabolism	Intronic	0.08	0.28	0.017	0.035	0.50	no
	rs16974794	Phase I metabolism	Intronic	0.08	0.26	0.020	0.035	0.50	no
	rs35950631	Phase I metabolism	Intronic	0.08	0.26	0.020	0.035	0.50	no
CYP2C9	rs28371685	Phase I metabolism	Coding synonymous	0.01	0.04	0.037	0.048	0.87	no
CYP2D6	rs1080984	Phase I metabolism	Near gene	0.03	0.12	0.012	0.012	0.50	no
	rs1080986	Phase I metabolism	Near gene	0.03	0.12	0.012	0.012	0.50	no
	rs1081003	Phase I metabolism	Coding synonymous	0.17	0.35	0.006	0.017	0.50	no
	rs28371706	Phase I metabolism	Coding nonsynonymous	0.05	0.17	0.013	0.021	1.00	yes
DRD2	rs11214608	Pharmacodynamic	Intronic	0.32	0.31	0.017	0.027	0.50	no
	rs1554929	Pharmacodynamic	Intergenic	0.27	0.29	0.019	0.028	0.50	no
	rs4587762	Pharmacodynamic	Intronic	0.32	0.30	0.018	0.028	0.50	no
	rs11214605	Pharmacodynamic	Intronic	0.31	0.30	0.019	0.030	0.50	no
	rs12798900	Pharmacodynamic	Intronic	0.31	0.30	0.019	0.030	0.50	no
	rs6274	Pharmacodynamic	3' UTR	0.04	0.14	0.032	0.043	0.50	no
	rs2234690	Pharmacodynamic	Intronic	0.33	0.27	0.030	0.047	0.50	no
	rs2587548	Pharmacodynamic	Intronic	0.33	0.27	0.030	0.047	0.50	no
F5	rs6020	Pharmacodynamic	Coding nonsynonymous	0.26	0.37	0.006	0.018	0.53	no
GSTT1	rs5996657	Phase II metabolism	Near gene	0.07	0.28	0.005	0.008	0.50	no
	rs6004035	Phase II metabolism	Near gene	0.07	0.27	0.009	0.015	0.50	no
	rs140313	Phase II metabolism	Intronic	0.37	0.28	0.026	0.039	0.50	no
	rs2739345	Phase II metabolism	Near gene	0.38	0.28	0.027	0.040	0.50	no
P2RY1	rs701265	Pharmacodynamic	Coding synonymous	0.34	0.27	0.030	0.045	0.63	yes
VKORC1	rs7200749	Pharmacodynamic	Coding nonsynonymous	0.06	0.21	0.013	0.022	0.60	no
	rs17878544	Pharmacodynamic	Near gene	0.10	0.30	0.016	0.037	0.50	no
	rs9923231	Pharmacodynamic	Near gene	0.47	0.39	0.003	0.004	0.50	yes

Abbreviations: SNV,single nucleotide variant; UTR, untranslated region; MAF, minor allele frequency.

[a]P-values are derived from the genome-wide empirical distribution of interpopulation F_{ST}.

[b]P-values are derived from the empirical distribution of interpopulation F_{ST} computed for SNVs with the same functional annotation than the SNV of interest.

Table 3.3 | Results of selection tests for the 13 VIP genes showing an atypical pattern of genetic differentiation. Four tests of positive selection (XP-EHH, XP-CLR, iHS and Tajima's *D*) were carried out at each gene locus in at least one of three population samples from the 1000 Genomes project (YRI, CEU and CHB). Statistical significance was assessed using genome-wide distributions of the test statistics, as previously described (Pybus *et al.* 2014). For each test, the percentage of scores significant at the 0.05 threshold is indicated and the number of significant scores is shown in parentheses. When the number of scores for a given test and gene region is above 4, the numbers are in bold. We considered the presence of a signal of positive selection in a given gene if (a) at least 10% of the computed scores were significant at the 0.05 threshold (with a minimum of four significant scores) for at least one selection test or if (b) at least 5% of the computed scores were significant at the 0.05 threshold (with a minimum of four significant scores) for at least two selection tests.

Gene	Population[a]	XP-EHH A[b]	XP-EHH B[c]	XP-CLR A[b]	XP-CLR B[c]	iHS	Tajima's *D*	Signal of positive selection
ADH1A	YRI	**28.5 (35)**	-	-	-	**16.9 (13)**	-	yes
	CHB	**96.7 (59)**	-	22.2 (2)	-	-	-	yes
ADH1B	YRI	-	**33.1 (40)**	-	-	**9.4 (8)**	-	yes
	CHB	**72.4 (55)**	**8.3 (10)**	**44.4 (4)**	**77.7 (7)**	-	-	yes
AHR	YRI	0.9 (2)	-	8.0 (2)	8.0 (2)	2.3 (3)	**76.5 (13)**	yes
	CEU	-	-	-	-	3.4 (2)	-	no
	CHB	-	-	-	-	1.8 (1)	-	no
CYP2B6	YRI	**5.1 (14)**	**7.0 (20)**	-	-	**2.7 (4)**	-	yes
	CEU	-	-	-	-	2.9 (3)	-	no
CYP2C9	YRI	**19.4 (6)**	-	-	-	-	-	yes
CYP2D6	YRI	1.8 (1)	-	-	33.3 (1)	4.5 (2)	-	no
	CHB	-	-	-	-	-	-	no
CYP3A4	CEU	**40.4 (36)**	**29.8 (48)**	13.3 (2)	**46.7 (7)**	7.9 (3)	**100.0 (10)**	yes
	CHB	-	-	-	-	4.0 (1)	**40.0 (4)**	yes
DRD2	YRI	**8.8 (47)**	**27.1 (140)**	-	-	**8.3 (29)**	-	yes
	CEU	**9.7 (38)**	-	5.3 (2)	5.3 (2)	**7.4 (16)**	-	yes
F5	YRI	**13.7 (31)**	1.3 (3)	4.0 (1)		**5.6 (27)**		yes
	CHB	**3.7 (18)**	**0.8 (5)**	**33.3 (13)**	2.6 (1)	0.7 (2)	-	yes
GSTT1	YRI	-	-	33.3 (2)	-	-	-	no
	CEU	-	-	**66.7 (4)**	**83.3 (5)**	-	-	yes
	CHB	-	-	-	-	-	-	no
P2RY1	CEU	-	**80 (44)**	-	-	-	-	yes
	CHB	-	**73.3 (44)**	-	-	-	-	yes
VDR	CEU	-	-	5.9 (2)	5.9 (2)	1.5 (3)	-	no
	CHB	-	**9.6 (55)**	-	5.9 (2)	**2.2 (4)**	-	no
VKORC1	YRI	**29.7 (11)**	-	-	-	-	-	yes
	CHB	**100.0 (27)**	**100.0 (38)**	100.0 (3)	66.7 (2)	-	100.0 (2)	yes

[a] YRI: Yoruba in Ibadan, Nigeria ; CEU : Utah residents with Northern and Western European ancestry; CHB: Han Chinese in Beijing, China.

[b] The CEU sample was used as a reference for YRI and CHB, and the CHB as a reference for CEU.

[c] The YRI sample was used as a reference for CEU and CHB, and the CHB as a reference for YRI.

2.3 Discussion

Identifier les facteurs génétiques impliqués dans la variabilité de réponse aux médicaments est un des enjeux majeurs de la pharmacogénétique. Pour ce faire, la démarche classique consiste à rechercher à partir de données sur des individus ayant reçu un même médicament les facteurs génétiques qui différencient les bons et mauvais répondeurs. Cette démarche est également celle des études cas-témoins, utilisées pour identifier les facteurs génétiques impliqués dans les maladies complexes. Elle présente le désavantage de nécessiter de gros échantillons d'individus exposés au médicament d'intérêt, chez qui les réponses thérapeutiques observées sont différentes. Dans certains cas de réactions adverses aux médicaments, l'identification de la molécule impliquée n'est pas évidente étant donné que les patients prennent souvent plusieurs médicaments en même temps. Une autre possibilité pour étudier le déterminisme génétique de la réponse aux médicaments consiste à réaliser des études de génétique des populations en comparant les distributions des allèles dans des pharmacogènes entre populations humaines à partir de panels d'individus représentatifs de ces populations. Ainsi, il est possible d'identifier des variants génétiques candidats qui pourront ensuite être étudiés en relation avec les médicaments pour lesquels on observe le plus de différences de réponse entre les populations humaines concernées. Cette démarche présente l'avantage d'être plus générique que la précédente puisqu'il n'est pas nécessaire de disposer d'une information sur l'exposition aux médicaments des individus des panels de diversité. C'est cette dernière démarche que nous avons utilisée dans notre étude en tirant parti de la récente disponibilité des données de séquence du Projet 1000 Génomes pour fournir la première analyse exhaustive de la différenciation géographique des 45 gènes majeurs intervenant dans la réponse aux médicaments, au sein de 14 populations humaines réparties en différents endroits du monde.

Notre analyse des profils de différenciation génétique des populations, estimés grâce à l'indice F_{ST}, a révélé une fraction élevée de variants très différenciés, en particulier pour les variants clés déjà connus pour jouer un rôle important en pharmacogénétique (Figure 3.2). Une différenciation géographique extrême pour ces variants peut se traduire par des différences

significatives de réponse aux médicaments entre les populations humaines, aussi bien en terme d'efficacité que de toxicité (Li et al., 2011).
C'est par exemple le cas du variant rs1229984, définissant l'allèle *ADH1B**2 codant pour la sous-unité ß2 de l'enzyme ADH1B. Un effet protecteur de cet allèle contre l'alcoolisme a été montré dans différentes populations humaines (Edenberg, 2000; Luo et al., 2006; Macgregor et al., 2009; Zuccolo et al., 2009). La susceptibilité aux pathologies liées à l'alcool est beaucoup plus faible en Asie que dans le reste du monde (Hasin and Grant, 2004). Cette différence peut s'expliquer en partie par une variation du taux de métabolisation de l'alcool entre les populations humaines, notamment liée à la plus grande prévalence de l'allèle dérivé de rs1229984 dans les populations asiatiques (Gemma et al., 2006).
Un autre exemple est, nous l'avons vu en détail dans la partie 2 de cette thèse, le variant rs9923231, en déséquilibre de liaison complet avec le variant rs9934438, qui confère un phénotype de sensibilité augmentée aux AVK (D'Andrea et al., 2005), médicaments pour lesquels on observe une grande variabilité inter-populationnelle de réponse (Schelleman et al., 2008).
Un autre exemple encore est le variant rs28371706 définissant l'allèle *CYP2D6*17* qui cause une diminution de l'activité enzymatique du CYP2D6 (Oscarson et al., 1997). Cet allèle a une distribution très hétérogène dans les populations humaines, étant principalement retrouvé dans les populations africaines (Bradford and Kirlin, 1998; Masimirembwa and Hasler, 1997). Il est donc susceptible de conduire à des différences de réponse entre africains et non africains pour de très nombreux médicaments métabolisés par le CYP2D6, notamment ceux utilisés en psychiatrie (neuroleptiques, antidépresseurs) et en cardiologie (bêta-bloquants) (Zanger and Schwab, 2013). En effet, une plus mauvaise réponse aux bêta-bloquants a été observée chez des individus d'origine africaine, en comparaison avec des individus d'origine européenne (Muszkat, 2007).

De manière intéressante, nous avons observé que la proportion de variants très différenciés au niveau mondial était supérieure pour les gènes de la pharmacodynamie que pour ceux de la pharmacocinétique (Figure 3.3). Il est possible que ce résultat soit le reflet de l'impact différentiel de la sélection

naturelle entre ces catégories de gènes. En effet, les gènes de la pharmacodynamie sont souvent des récepteurs activés par des ligands spécifiques tandis que ceux de la pharmacocinétique peuvent être impliqués dans la biotransformation et le transport d'un grand nombre de molécules. Ces derniers gènes peuvent donc faire l'objet de pressions de sélection de différentes natures pouvant agir de manière antagoniste, favorisant un variant ou un autre selon la molécule considérée. Il est donc peu probable qu'un même variant soit globalement favorisé dans certaines populations, conduisant à des valeurs élevées de F_{ST}. En revanche, les gènes ayant un spectre de substrats très limité, comme *VKORC1* et la vitamine K, peuvent plus facilement faire l'objet d'une sélection ciblée sur un variant donné. Cette hypothèse, purement spéculative, nécessite d'être confirmée en évaluant l'impact de la sélection naturelle sur les différentes catégories de pharmacogènes à une plus grande échelle.

Parmi l'ensemble des variants présentant des profils de différenciation extrême, nous en avons sélectionné 35 fonctionnels ou potentiellement fonctionnels qui pourraient jouer un rôle dans la réponse aux médicaments. Ces 35 variants sont situés dans 13 gènes différents (Tableau 3.2) et parmi eux, 27 sont des variants nouveaux n'ayant jamais été associés à la réponse aux médicaments (Tableau 3.4).

A l'instar de différentes études analysant la différenciation pharmacogénétique des populations humaines, nous avons observé que dans la majorité des profils de différenciation extrême obtenus pour ces 35 variants, l'Afrique était la région qui se distinguait le plus des autres régions (Figures 3.4 et 3.5). Cette observation souligne l'importance des différences pharmacogénétiques retrouvées entre les populations africaines et les populations non africaines et des conséquences que celles-ci peuvent avoir sur un plan pharmacologique et toxicologique. En conséquence, il apparaît important d'insister sur la nécessité d'inclure des populations africaines, plus diversifiées génétiquement que les autres populations du monde, dans les études de pharmacogénétique (Aminkeng et al., 2014; Dandara et al., 2014; Delser and Fuselli, 2013; Masimirembwa and Hasler, 1997; Ramos et al., 2013).

Pour chacun des 13 gènes, nous avons recherché des empreintes génomiques de la sélection positive dans les régions géographiques où l'allèle dérivé du variant très différencié est le plus fréquent (en excluant les populations américaines qui sont des populations mélangées, plus difficiles à analyser). Pour rechercher ces empreintes, nous avons utilisé plusieurs tests de sélection qui permettent de détecter des évènements de sélection ayant des caractéristiques différentes. L'iHS et l'XP-EHH sont en effet capables de détecter des balayages sélectifs récents, complets (XP-EHH) ou partiels (iHS) (Pickrell et al., 2009), tandis que l'XP-CLR et le *D* de Tajima peuvent détecter des évènements sélectifs plus anciens (Chen et al., 2010; Sabeti et al., 2006). En combinant ces différents tests, nous pouvons espérer avoir une meilleure puissance pour détecter un balayage sélectif, même si cela peut avoir un coût en terme de tests multiples que nous n'avons pas discuté mais qui mériterait néanmoins d'être évalué.

Comme le montre la Figure S6 (*cf.* Annexe 2), un signal de sélection positive a été identifié pour 11 des 13 gènes (soit pour plus de 80 % des gènes testés), suggérant que la sélection naturelle est une force évolutive majeure dans la répartition géographique des variations génétiques au sein des pharmacogènes. Le Tableau 3.4 ci-dessous fournit pour chacun des 13 gènes étudiés la liste des variants d'intérêt identifiés dans notre étude, les associations connues pour ces variants avec les phénotypes de la réponse aux médicaments ainsi que les signatures de sélection identifiées dans notre étude et celles déjà reportées dans la littérature. Nous retrouvons tous les signaux de sélection déjà identifiés : pour les gènes *ADH1A* et *ADH1B* en Asie de l'Est, pour *CYP3A4* dans les populations non africaines, pour *DRD2* en Afrique et pour *VKORC1* en Asie de l'Est (Tableaux 3.3 et 3.4). En revanche, nous ne mettons pas en évidence le signal de sélection déjà décrit dans la littérature pour le gène *DRD2* en Asie de l'Est (Lao et al., 2007). Ceci peut s'expliquer par notre choix d'une stratégie d'étude de la sélection focalisée sur les variants dont la différenciation génétique globale est atypique et qui ne cherche donc pas à détecter un signal sélectif en aveugle dans chacune des trois populations testées (CEU, CHB et YRI) pour chacun de ces 13 gènes. En effet, la fréquence faible des allèles dérivés en Asie des variant très

différenciés de *DRD2* (Figure 3.5) ne nous a pas conduit à explorer la sélection dans la population asiatique CHB pour ce gène. Cette constatation illustre les limites d'une approche basée sur un seul ou quelques variants dans le gène pour étudier la sélection.

En plus de ces signaux déjà décrits, nous détectons de la sélection positive pour la première fois dans six gènes : *AHR*, *CYP2B6*, *CYP2C9*, *F5*, *GSTT1* et *P2RY1*. Nous ne pouvons pas être sûrs que ces gènes représentent vraiment la cible de la sélection car, sans étudier en détail les régions génomiques situées autour de ces gènes comme nous l'avions fait dans l'étude de *VKORC1* (*cf.* partie 2 de cette thèse), nous ne pouvons pas exclure la possibilité d'un phénomène d'auto-stop génétique. Cependant, pour certains de ces gènes nous pouvons émettre des hypothèses sur l'action possible de la sélection. Par exemple, pour le gène *GSTT1* une étude réalisée dans une population iranienne a suggéré qu'il pourrait favoriser la survie des individus exposés à des composés chimiques toxiques présents dans un gaz naturel (Zendeh-Boodi and Saadat, 2008). Par ailleurs, le signal de sélection que nous identifions sur le gène *AHR* dans la population africaine YRI pourrait être relié au rôle crucial du récepteur aux hydrocarbures aromatiques (AhR) codé par ce gène dans les processus de détoxification de l'organisme contre les polluants environnementaux. Parmi les ligands de l'AhR figurent des composés très toxiques tels que les hydrocarbures aromatiques polycycliques comme le benzopyrène, les biphényls chlorés plans et les hydrocarbures aromatiques halogénés comme les dioxines (Denison et al., 2002). L'AhR reconnaît la présence des xénobiotiques dans le cytoplasme. Leur fixation à ce récepteur entraîne leur translocation du compartiment cytoplasmique vers le noyau et induit l'expression des gènes impliqués dans le métabolisme et l'élimination des composés étrangers, dont certains membres de la famille des cytochromes P450 comme *CYP1A1, CYP1A2, CYP1B1*, ainsi que d'autres enzymes de détoxification comme la GST, *NQO1, GSTA2, UGT1A1, UGT1A6* et *Nrf2 (NFE2L2*) (Kawajiri and Fujii-Kuriyama, 2007). Le rôle de l'AhR dans la régulation des gènes du métabolisme des xénobiotiques est considéré comme étant une fonction adaptative (Köhle and Bock, 2007) .

Il faut cependant rester prudent dans ces hypothèses, car l'identification des molécules à l'origine des pressions de sélection est souvent très difficile. En effet, il est même possible que certains pharmacogènes impliqués dans la biotransformation d'un large éventail de substrats aient subi plusieurs pressions sélectives dirigées sur un substrat ou une catégorie de substrats particuliers conférant un avantage dans un environnement donné. C'est le cas par exemple, comme indiqué dans le Tableau 3.4, du gène *ADH1B,* impliqué dans la biotransformation des xénobiotiques, particulièrement de l'alcool. Il a été proposé que la sélection positive sur le gène *ADH1B* en Asie de l'Est pourrait avoir favorisé les individus ayant un métabolisme de l'alcool augmenté au moment de la hausse de la consommation d'alcool conjointe à l'apparition de la culture du riz (Peng et al., 2010). D'un autre côté, ce gène peut également avoir subi l'action de la sélection en lien avec sa capacité de détoxification, qui aurait été avantageuse pour lutter contre les pathogènes ou les toxines alimentaires (Goldman and Enoch, 1990). Il en est de même pour le gène *CYP3A4*, impliqué dans le métabolisme de plus de la moitié des médicaments employés aujourd'hui. Plusieurs hypothèses adaptatives ont été proposées pour expliquer la sélection positive détectée dans ce gène dans les populations non africaines. Ainsi, il est possible que ce gène ait conféré un avantage sélectif en ayant un rôle protecteur contre le rachitisme, *via* son implication dans le métabolisme de la vitamine D (Schirmer et al., 2006). Par ailleurs, des phénomènes adaptatifs en réponse à une augmentation de toxines environnementales tels que les acides biliaires pourraient être à l'origine de la sélection (Krasowski et al., 2005). Cette hypothèse a notamment été confirmée par le fait que l'acide lithocholique, un acide biliaire, entraîne une augmentation drastique de la synthèse de CYP3A4 spécifiquement chez l'homme, suggérant qu'il est responsable de l'évolution rapide du spectre de substrats du CYP3A4 dans la lignée humaine (Kumar et al., 2009). Le changement rapide de séquence du CYP3A4 représenterait un nouveau mécanisme de défense contre la cholestase (urgence diagnostique, engendrée par une obstruction de la bile dans les voies biliaires), en réponse à une augmentation de l'apport en stéroïdes dans l'alimentation chez l'homme, notamment les acides biliaires (Kumar et al., 2009).

En définitive, notre étude identifie une liste de nouveaux variants d'intérêt potentiel en pharmacogénétique pouvant expliquer une part de la variabilité inter-populationnelle, mais aussi interindividuelle, de réponse aux médicaments. Ces variants méritent d'être étudiés de manière plus approfondie par la mise en œuvre d'études fonctionnelles permettant d'évaluer leur impact sur la fonction ou l'expression de la protéine produite et de préciser leur rôle dans la réponse aux médicaments *via* un mécanisme pharmacocinétique ou pharmacodynamique. Leur inclusion dans des études d'associations pharmacogénétiques pourra permettre de déterminer leur relation avec les phénotypes de la réponse aux médicaments dans les différentes populations humaines.

De plus, notre étude démontre l'importance de la sélection naturelle sur la différenciation génétique des pharmacogènes. Il est possible que d'autres gènes inclus dans notre étude aient également subi l'action de la sélection naturelle. En effet, nous n'avons recherché une signature génomique de sélection que dans les 13 gènes possédant au moins un variant clé ou un variant potentiellement délétère très différencié entre les populations humaines. Par ailleurs, il est possible que d'autres types de sélection aient agi sur les pharmacogènes. En particulier, bien que de façon cohérente avec les résultats de la littérature (Jorge et al., 1999), nous ne détectons pas de signature génomique de sélection pour le gène *CYP2D6*, il a été proposé que ce gène soit sous l'action de la sélection naturelle dans les populations d'Afrique du Nord Est, qui favoriserait le portage de multiples copies du gène *CYP2D6*. Cette hypothèse, en rapport avec l'alimentation, a été avancée pour expliquer les différences dans l'activité du CYP2D6 retrouvée entre les éthiopiens natifs d'Ethiopie et les éthiopiens vivant en Suède (Ingelman-Sundberg, 2005). Des alcaloïdes présents dans l'alimentation éthiopienne, inhibant l'activité du CYP2D6, pourraient représenter un facteur de sélection : en période de famine, les individus capables de détoxifier les toxines de plantes à un plus large degré auraient été favorisés, augmentant ainsi le nombre de plantes pouvant potentiellement servir d'aliments. Les méthodes mises en œuvre dans cette étude ne sont clairement pas adaptées pour mettre en évidence ce type de sélection portant sur le nombre de copies d'un gène.

Tableau 3.4 | Résumé du rôle des 13 pharmacogènes présentant un variant d'intérêt en pharmacogénétique (variant déjà connu ou variant candidat) très différencié entre les populations humaines et des résultats de sélection obtenus dans notre étude.

Gène VIP	Rôle	Lien avec la réponse aux médicaments et les maladies	Présence d'un variant clé possédant un F_{ST} significatif dans 1KG	Identification de nouveaux variants candidats	Associations connues des variants candidats avec la réponse aux médicaments	Pression de sélection décrite pour ce gène dans la littérature	Hypothèse possible de pression sélective	Détection de la sélection positive dans les populations 1KG testées
ADH1A (chr 4q22-23)	Code pour la sous-unité alpha de l'alcool déshydrogénase de classe 1, exprimée pendant la vie fœtale.	Rôle important dans les processus de détoxification de l'organisme et dans le métabolisme de l'alcool (Edenberg, 2000). Associé avec le risque de dépendance aux médicaments de type opiacés et à des dérèglements psychiques (Luo et al., 2007)	- rs975833	- rs6854980 - rs28364313 - rs3819197	Non	Le locus *ADH* sort dans un scan de sélection positive génome entier (Voight et al., 2006).	La sélection naturelle sur les enzymes ADH pourrait être liée aux toxines environnementales, retrouvées dans des changements de l'alimentation ou les maladies infectieuses (Goldman et al., 1990).	YRI : oui CEU : pas testé CHB : oui
ADH1B (chr 4q22-23)	Code pour la sous-unité bêta de l'alcool déshydrogénase de classe 1, exprimée pendant la vie adulte.	Rôle important dans les processus de détoxification de l'organisme, dans la clairance des médicaments (Crabb et al., 2004), et dans le métabolisme de l'alcool (Edenberg, 2000)	- rs1229984 G>A (Arg48His) (allèle *ADH1B*2*) Allèle protecteur contre l'alcoolisme dans des populations de différentes origines ethniques (Celorrio et al., 2011; Macgregor et al., 2009; Luo et al., 2006; Zuccolo et al., 2009; Edenberg, 2000). - rs2066702 (Arg370Cys) (allèle *ADH1B*3*) Associé à la dépendance aux médicaments (Luo et al., 2007) et à l'alcool (Luo et al., 2006; Zuo et al., 2013) De manière intéressante, la dépendance à l'alcool varie selon l'origine ethnique (Gemma et al., 2006; Hasin and Grant, 2004), et ces variants sont fortement différenciés entre les populations humaines (Osier et al., 2002).	- rs2070898 - rs3811801	Non	Le locus *ADH* sort dans un scan de sélection positive génome entier (Voight et al., 2006). Un autre scan de sélection positive génome entier détecte le gène *ADH1B* (Barreiro et al., 2008). Des études gènes candidats ont détecté de la sélection positive pour *ADH1B* en Asie de l'Est (Han et al., 2007; Li et al., 2008; Xue et al., 2009; Peng et al., 2010).	La sélection pourrait être en lien avec : - Les modes de vie (Li et al., 2008) - L'apparition de la culture du riz aurait été suivie d'une augmentation de la consommation d'alcool. L'allèle *ADH1B*2* aurait été avantagé car il permet une élimination 100 fois plus rapide de l'alcool que l'allèle *ADH1B*1* (Peng et al., 2010). La sélection n'est pas retrouvée dans les populations tibétaines dont l'alimentation n'est pas orientée vers l'alimentation animale, confirmant cette hypothèse d'adaptation sélective liée à l'agriculture (Lu et al., 2012). - Il semble que le variant rs3811801 soit la cible directe de la sélection car il permet de potentialiser l'effet de l'allèle *ADH1B*2* (Li et al., 2008).	YRI : oui CEU : pas testé CHB : oui

Tableau 3.4 (suite)

Gène VIP	Rôle	Lien avec la réponse aux médicaments et les maladies	Présence d'un variant clé possédant un F_{ST} significatif dans 1KG	Identification de nouveaux variants candidats	Associations connues des variants candidats avec la réponse aux médicaments	Pression de sélection décrite pour ce gène dans la littérature	Hypothèse possible de pression sélective	Détection de la sélection positive dans les populations 1KG testées
AHR (chr 7p15)	Code pour le récepteur aux hydrocarbures aromatiques (AhR), un facteur de transcription activé par la fixation de ligands (xénobiotiques). Induit l'expression de gènes impliqués dans le métabolisme et l'élimination des composés étrangers.	Les gènes cibles de l'AhR comptent des gènes impliqués dans la détoxification des xénobiotiques et le métabolisme des médicaments : *CYP1A1*, *CYP1A2*, *CYP1B1*, *GST*, *NQO1*, *GSTA2*, *UGT1A1*, *UGT1A6*, et *NFE2L2*.	Non	- rs1557737 (G>A) - rs35618592 (G>T)	Non	Sélection purificatrice chez la souris, le rat et l'homme (Thomas et al., 2002).	En lien avec le rôle crucial d'AhR dans les processus de détoxification. Sans doute une réponse adaptative aux polluants environnementaux.	YRI : oui CEU : non CHB : non
CYP2B6 (chr 19q13.2)	Métabolisme hépatique de phase I	Métabolise 4 % des médicaments les plus prescrits (Zanger et al., 2008), dont : - anticancéreux (cyclophosphamide, fosphamide), - antirétroviraux (éfavirenz, névirapine), - antidépresseurs (bupropion), - antipaludiques (artémisinine), - nicotine, - prasugel, - anesthésiques (propofol), - opioïdes synthétiques.	Non	- rs12721652 - rs28739581 - rs35950631 - rs16974794 - rs28399501	- rs12721652 : impliqué dans la variabilité interindividuelle de réponse au bupropion (un antidépresseur couramment utilisé dans le traitement de l'addiction à la nicotine) (Hesse et al., 2004) - rs28739581: non - rs35950631: non - rs16974794 : non - rs28399501 : non	Non	-	YRI : oui CEU : non CHB : pas testé
CYP2C9 (chr 10q24)	Métabolisme hépatique de phase I	Métabolise 10 à 15 % des molécules utilisées en thérapeutique, dont: - anti-inflammatoires non stéroïdiens (AINS), - anticoagulants oraux, - hypoglycémiants oraux, - antiépileptiques, - antagonistes du récepteur de l'angiotensine II, - fluvastatine (Muszkat, 2007).	Non	rs28371685 C>T, Arg335Trp	Affecte la biodisposition de l'aspirine (Agúndez et al., 2009).	Non	-	YRI : oui CEU : pas testé CHB : pas testé

Tableau 3.4 (suite)

Gène VIP	Rôle	Lien avec la réponse aux médicaments et les maladies	Présence d'un variant clé possédant un F_{ST} significatif dans 1KG	Identification de nouveaux variants candidats	Associations connues des variants candidats avec la réponse aux médicaments	Pression de sélection décrite pour ce gène dans la littérature	Hypothèse possible de pression sélective	Détection de la sélection positive dans les populations 1KG testées
CYP2D6 (chr 22p13.1)	Métabolisme hépatique de phase I	Métabolise 25 % des médicaments utilisés en clinique (Wang et al., 2009), dont : - Les neuroleptiques, - Les antidépresseurs, - Les opiacés faibles, - Certains antiarythmiques (Ingelman-Sundberg, 2005). Différents phénotypes métaboliques selon le niveau d'activité enzymatique du CYP2D6 : métaboliseurs ultrarapides, normaux, intermédiaires et lents. Induisent des différences de réponse aux médicaments entre les individus (toxicité, inefficacité).	- rs61736512 (V136M) - rs59421388 (V287M) - rs28371706 (T107L)	- rs1080984 - rs1080986 - rs1081003	Non	- Oui chez les animaux (Schenekar et al., 2011). - Non chez l'homme (Jorge et al., 1999).	- Pour les animaux : lien avec l'alimentation ou la capacité de détoxifier les alcaloïdes. - Chez l'homme : une hypothèse de sélection a été proposée pour expliquer les différences d'activité du CYP2D6, en rapport avec l'alimentation. Les individus ayant une bonne capacité à détoxifier les toxines présentes dans les plantes auraient été avantagés en période de famine (Ingelman-Sundberg, 2005).	YRI : non CEU : non CHB : pas testé
CYP3A4 (chr 7q21.1)	Métabolisme hépatique de phase I	Métabolise 50-60 % des médicaments utilisés en clinique, dont : - Antibiotiques (macrolides) - Antiarythmiques - Certaines benzodiazépines - Immunomodulateurs - Inhibiteurs de protéase - Antihistaminiques - Bloqueurs des canaux calciques - Statines - Paracétamol - Hormones stéroïdiennes - Ciclosporine - … Grande variabilité d'activité métabolique.	- rs2740574 c.-392G>A (allèle *CYP3A4*1B*) Associé à différents types de cancers (Zhou et al., 2013).	Non	-	Oui : - Sélection positive dans les populations non africaines (Thompson et al., 2004; Schirmer et al., 2006) - Sélection positive sur la séquence protéique chez l'homme, alors qu'on retrouve de la sélection purificatrice chez les primates (Qiu et al., 2008) - Les sites de liaison du CYP3A4 sont sous sélection dans les gènes *PXR* et *CAR* (Krasowski et al., 2005) - Sélection positive dans les populations africaines, européennes, asiatiques (Chen et al., 2009) - Sélection positive en Asie (Zhou et al., 2011).	La sélection pourrait être en lien avec : - le rôle de CYP3A4 dans le métabolisme de la vitamine D. Le rachitisme pourrait avoir été la pression sélective en faveur de l'éradication de l'allèle ancestral G du SNP rs2740574 (Schirmer et al., 2006). - l'apparition de la cuisson des aliments chez les humains, qui aurait entrainé l'augmentation du spectre de substrats aujourd'hui métabolisés par le CYP3A4 (Qiu et al., 2008), notamment les sels biliaires (Krasowski et al., 2005). L'acide lthocholique semble être responsable de l'augmentation de la synthèse du CYP3A4, pour lutter contre la cholestase au moment de l'apport augmenté en stéroïdes dans l'alimentation (Kumar et al., 2009). - Les variations de l'environnement chimique (Chen et al., 2009).	YRI : pas testé CEU : oui CHB : oui

Tableau 3.4 (suite)

Gène VIP	Rôle	Lien avec la réponse aux médicaments et les maladies	Présence d'un variant clé possédant un F_{ST} significatif dans 1KG	Identification de nouveaux variants candidats	Associations connues des variants candidats avec la réponse aux médicaments	Pression de sélection décrite pour ce gène dans la littérature	Hypothèse possible de pression sélective	Détection de la sélection positive dans les populations 1KG testées
DRD2 (chr 11q23)	Impliqué dans la formation des récepteurs à la dopamine	La dopamine est un médicament utilisé en psychiatrie (dans le traitement de la dépression) et dans le traitement de la maladie de Parkinson.	- rs6277 G>A (Pro319Pro)	- rs11214608 - rs6274 - rs4587762 - rs11214605 - rs12798900 - rs2234690 - rs2587548 - rs1554929	- rs1554929 : associé avec la migraine (Corominas et al., 2009) et un phénotype relié à l'alcool (Meyers et al., 2013) - Aucune association connue pour les sept autres variants.	Sélection positive en Asie et en Afrique (Lao et al., 2007)	La sélection détectée pourrait être en lien avec l'association de DRD2 avec la voie métabolique de la mélanine, car les phénotypes de couleur de la peau sont susceptibles d'être soumis à des phénomènes d'adaptation dans les populations humaines (Lao et al., 2007).	YRI : oui CEU : oui CHB : pas testé
F5 (chr 1q23)	Code pour le facteur 5 de la coagulation	Le variant FVL (Facteur 5 de Leiden) a été évalué comme étant un facteur de risque de plusieurs maladies (occlusion intestinale, mortalité dans les septicémies, syndrome de Budd-Chiari, complications obstétriques, infarctus cérébral et la porencéphalie) en plus de la thrombophilie et parahémophilie (Whirl-Carrillo et al., 2012).	Non	- rs6020 (C>T Arg513Lys)	Non	Non	-	YRI : oui CEU : pas testé CHB : oui
GSTT1 (chr 22Q11.23)	Code pour l'enzyme glutathion S-transférase (GST) 1	La délétion complète du gène *GSTT1* a été associée à la leucémie myéloïde aigüe (Mossalam 2006), et représente un facteur de risque pour le DILI *Drug-induced liver injury* consécutive à un traitement par isoniazide et troglitazone (Daly, 2010).	Non	- rs2739345 - rs140313 - rs5996657 - rs6004035	Non	Pas de signal détecté dans deux études (Lordelo et al., 2012; Polimanti et al., 2011)	Hypothèse d'une sélection en cours dans une population iranienne, qui expliquerait la diminution de la fréquence du génotype nul entre deux générations, en rapport avec l'exposition à un gaz naturel riche en sulfure d'hydrogène (Zendeh-Boodi and Saadat, 2008).	YRI : non CEU : oui CHB : non

Tableau 3.4 (suite)

Gène VIP	Rôle	Lien avec la réponse aux médicaments et les maladies	Présence d'un variant clé possédant un F_{ST} significatif dans 1KG	Identification de nouveaux variants candidats	Associations connues des variants candidats avec la réponse aux médicaments	Pression de sélection décrite pour ce gène dans la littérature	Hypothèse possible de pression sélective	Détection de la sélection positive dans les populations 1KG testées
P2RY1 (chr 3q25.2)	Code pour le récepteur purinergique plaquettaire P2Y , récepteur couplé à une protéine G.	Associé avec la variabilité de l'aggrégation plaquettaire (Hetherington et al., 2005), impliqué de manière indirecte dans la réponse au clopidogrel, un antiaggrégant plaquettaire (Lev et al., 2007).	- rs701265	Non	-	Non	-	YRI : pas testé CEU : oui CHB : oui
VDR (chr 12q13.11)	Code pour le récepteur de la vitamine D, qui appartient à la superfamille des récepteurs nucléaires.	Module l'expression de gènes en se fixant au niveau de la séquence VDRE (*Vitamin D rich elements*), dont certains gènes ayant une fonction vitale (Poon et al., 2012). Entre 1800 et 2600 gènes sont ciblés par *VDR*. La vitamine D intervient dans le métabolisme phosphocalcique, et la minéralisation de l'os (DeLuca et al., 2004), ainsi que dans la différenciation et prolifération des cellules intestinales, rénales et immunitaires (Carlberg et al., 2013). Le polymorphisme génétique de *VDR* est impliqué dans le rachitisme, l'ostéoporose, l'arthrose, certains cancers, l'hypertrophie de la prostate, l'hyperparathyroïdisme,le diabète, la susceptibilité à certaines infections, le psoriasis, les coronopathies (Zmuda et al., 2000).	- rs11568820	Non	-	Pas de sélection positive détectée pour ce gène, qui semble être au contraire, sous une forte sélection purificatrice (Krasowski et al., 2005). En revanche, les nombreux ligands de *VDR* semblent être enrichis en signaux de sélection positive (Ramagopalan et al., 2010).	-	YRI : pas testé CEU : non CHB : non
VKORC1 (chr 16p11.2)	Code pour l'enzyme vitamine K oxydo-réductase, qui intervient dans la synthèse des protéines vitamine K dépendantes.	Représente la cible pharmacologiques des anticoagulants oraux de type AVK.	- rs9923231 (-1639G>A) - rs9934438 (1173 C>T)	- rs17878544 - rs7200749	- rs17878544 associé avec l'augmentation de dose requise d'acécoumarol dans une population caucasienne (Mitchell et al., 2011). - rs7200749 associé avec l'augmentation de dose requise de warfarine dans une population africaine (Mitchell et al., 2011).	De la sélection positive a été détectée en Asie de l'Est (Ross et al. 2010), mais une étude apporfondie de ce signal a conclu que ce gène pouvait ne pas être la cible directe de la sélection (Patillon et al., 2012)	La sélection pourrait être en rapport avec le métabolisme de la vitamine K, ou avec l'apparition d'une moécule anticoagulante de type AVK dans l'environnement (Patillon et al., 2012).	YRI : oui CEU : pas testé CHB : oui

Partie 4

Étude du gène *NAT2*

Chapitre 1

Généralités sur le polymorphisme d'acétylation

L'étude des pharmacogènes présentée dans la partie précédente comprenait l'ensemble des gènes autosomaux annotés comme étant des gènes majeurs de la réponse aux médicaments (*VIP genes*) dans la base de données PharmGKB. Cette liste est régulièrement actualisée par les membres du PGRN et elle inclut aujourd'hui, entre autres, le gène *NAT2* (*N*-acétyltransférase 2) qui présente un intérêt tout particulier en pharmacogénétique du fait de son importante variabilité fonctionnelle et de son implication dans le métabolisme d'un grand nombre de molécules thérapeutiques utilisées en clinique.

1. L'enzyme NAT2

L'arylamine *N*-acétyltransférase 2 (NAT2) est une enzyme de conjugaison intervenant dans la phase II du métabolisme des xénobiotiques. Elle est principalement exprimée dans le foie et l'intestin grêle. Elle catalyse le transfert d'un groupement acétyl sur les métabolites de type amines aromatiques et hétérocycliques, les rendant ainsi plus hydrophiles et plus faciles à éliminer dans les fluides biologiques. Elle participe ainsi à la biotransformation de nombreux xénobiotiques, incluant des molécules pré-cancérogènes présentes dans la fumée de cigarette, l'environnement, les pesticides et l'alimentation, et un certain nombre de médicaments utilisés en routine clinique. Citons par exemple l'hydralazine (antihypertenseur), les sulfamides antibactériens tels que la sulfadimidine ou la sulfapyridine, le procaïnamide (antiarythmique), les benzodiazépines telles que le clonazépam et le nitrazépam (anxiolytiques), l'isoniazide (antituberculeux) ou

encore la caféine (Weber and Hein, 1985). Si l'enzyme NAT2 contribue aux processus de détoxification de l'organisme en facilitant l'élimination de certaines molécules, elle peut aussi contribuer à la bioactivation de molécules pré-cancérogènes, qui requièrent une activation métabolique pour initier la carcinogénèse, comme c'est le cas pour des amines aromatiques dans le cancer de la vessie (Butcher et al., 2002; Hein, 2006).

2. Le polymorphisme d'acétylation

Un important polymorphisme génétique dans le gène codant pour NAT2 est à l'origine d'une grande variation de son activité enzymatique et explique, en partie, l'importante variabilité interindividuelle dans la capacité d'acétylation. Ce polymorphisme d'acétylation est l'un des premiers traits héréditaires affectant la réponse individuelle aux médicaments à avoir été découvert chez l'homme et il est probablement aujourd'hui l'un des mieux documentés. De fait, il occupe une place importante dans le domaine de la pharmacogénétique. Il peut avoir de nombreuses conséquences en clinique, notamment au niveau de la réponse pharmacologique et toxicologique aux médicaments. En outre, il représente un facteur de susceptibilité à certains cancers et autres maladies complexes.

2.1 Découverte

A l'instar des AVK dont l'introduction a conduit des années plus tard à l'identification du gène *VKORC1* et du variant fonctionnel rs9923231 conférant la sensibilité augmentée à ces molécules, la découverte du polymorphisme d'acétylation chez l'homme illustre bien l'approche expérimentale classique des études pharmacogénétiques : l'observation en premier lieu d'une variabilité interindividuelle de réponse à un traitement médicamenteux qui conduit, après investigation clinique, biochimique et génétique, à la découverte des gènes et des variants associés aux différents phénotypes de réponse aux médicaments observés.

C'est suite à l'introduction de l'isoniazide comme traitement de la tuberculose qu'a été décrite pour la première fois la variabilité

interindividuelle de la capacité d'acétylation (BONICKE and REIF, 1953). La survenue de réactions de toxicité (neuropathies périphériques) chez près de 40 % des patients traités avec cette molécule, du fait de l'accumulation de la forme non transformée de l'isoniazide, a révélé l'existence d'individus éliminant de façon moins efficace le médicament (HUGHES et al., 1954). Plusieurs études ont mis en évidence, comme illustré dans la Figure 4.1, une distribution bimodale du taux plasmatique d'isoniazide après administration d'une dose standard chez des patients traités par ce médicament, différenciant ainsi les acétyleurs rapides des acétyleurs lents (EVANS et al., 1960; MITCHELL and BELL, 1957). Dans les années suivantes, il est apparu que de nombreuses autres molécules, étaient également concernées par ce polymorphisme d'acétylation (Weber et Hein, 1985).

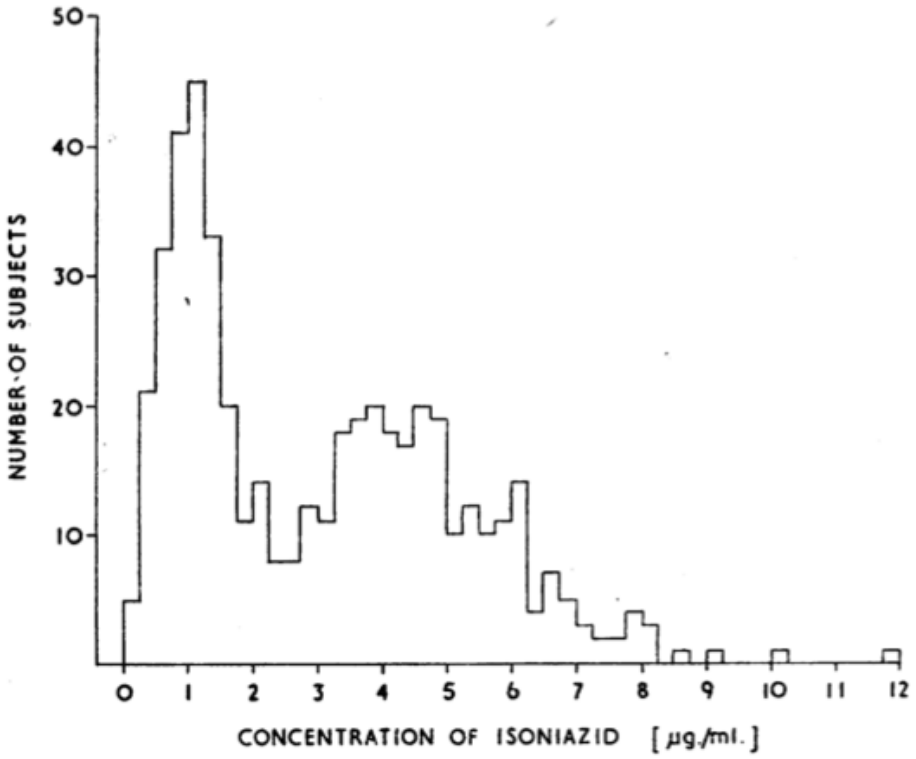


Figure 4.1 | Concentration plasmatique d'isoniazide mesurée six heures après l'administration d'une dose standard du médicament (10 mg/kg) chez 483 sujets. Tiré de (EVANS et al., 1960).

2.2 Conséquences cliniques

L'importante variabilité interindividuelle de la capacité d'acétylation peut se traduire par une modification de la réponse pharmacologique et toxicologique aux médicaments acétylés mais aussi par une susceptibilité différentielle à un certain nombre de maladies complexes.

Ainsi, les individus acétyleurs lents peuvent présenter un risque accru de réactions indésirables aux médicaments, dont la plus importante est l'hépatotoxicité induite par l'isoniazide (Huang et al., 2002; Leiro-Fernandez et al., 2011). D'autres effets toxiques des médicaments peuvent être observés, tels que des neuropathies périphériques induites par l'isoniazide, le lupus érythémateux disséminé induit par l'isoniazide, l'hydralazine, ou le procaïnamide, ou encore des phénomènes d'intolérance à la salicylazosulfapyridine (cyanose et hémolyse) ou à la phénelzine (somnolence et nausées) (Larrey et al., 1985; Weber and Hein, 1985). A l'inverse, une dose standard du médicament peut être inefficace chez les acétyleurs rapides. Par exemple ces sujets requièrent des doses de phénelzine, de dapsone ou encore d'hydralazine augmentées pour bénéficier d'un effet thérapeutique adéquat.

Par ailleurs, le polymorphisme d'acétylation a été associé à des différences de susceptibilité à certaines maladies complexes, notamment des cancers. L'association la plus convaincante concerne le cancer de la vessie pour lequel les individus acétyleurs lents seraient plus à risque (Hein, 2006; Sanderson et al., 2007; Selinski et al., 2013). Des études récentes ont mis en évidence que ces sujets auraient également une susceptibilité augmentée au cancer de la tête et du cou (Khlifi et al., 2013), au cancer du sein et au cancer de l'estomac (Fernandes et al., 2013). Une association a été retrouvée chez des sujets acétyleurs lents avec le gliome (Muthusamy et al., 2012). Les individus acétyleurs rapides auraient un risque plus élevé de développer un cancer colorectal (Liu et al., 2012a). Récemment, une augmentation de la susceptibilité à l'asthme chez les individus acétyleurs lents a également été rapportée (Wang et al., 2014).

2.3 Déterminants génétiques

La variabilité interindividuelle de la capacité d'acétylation *in vivo* est déterminée par des polymorphismes génétiques situés dans l'unique exon codant (exon 2) du gène *NAT2*. Cet exon de 870 pb est extrêmement polymorphe avec plus de 38 variations nucléotidiques décrites à ce jour (http://nat.mbg.duth.gr/). Parmi elles, sept polymorphismes majeurs se distinguent, dont les conséquences sur l'activité enzymatique de NAT2 sont bien caractérisées (Ferguson et al., 1994; Fretland et al., 2001; Hein et al., 1994; Leff et al., 1999). Quatre entraînent un changement d'acide aminé qui conduit à une diminution significative de l'activité d'acétylation (G191A, T341C, G590A et G857A), par altération de l'activité catalytique ou diminution de l'expression ou de la stabilité de l'enzyme. Les trois autres sont des mutations silencieuses (C282T, C481T) ou des substitutions non synonymes (A803G) qui n'altèrent pas le phénotype d'acétylation *in vivo*.

2.4 Corrélation génotype-phénotype

Traditionnellement, la détermination du statut d'acétyleur d'un individu est réalisée par l'utilisation de méthodes de phénotypage *in vivo*. Celles-ci consistent en l'évaluation de l'activité enzymatique de NAT2 par la mesure de la concentration d'un substrat test et de ses principaux métabolites, dans un échantillon urinaire ou sanguin, après administration d'une dose standard de ce substrat. La valeur du rapport métabolique mesuré permet de déterminer si un individu est acétyleur rapide ou acétyleur lent. Le phénotypage par la caféine est aujourd'hui considéré comme la méthode de choix pour mesurer l'activité de NAT2 *in vivo*. Certains auteurs ont décrit une distribution trimodale des rapports métaboliques permettant de discriminer en plus les acétyleurs intermédiaires des acétyleurs rapides (Grant et al., 1984; Kilbane et al., 1990; Pontes et al., 1993; Saruwatari et al., 2002; Tang et al., 1987, 1991).

Cependant, cette technique est de plus en plus remplacée par le génotypage des principaux polymorphismes fonctionnels de *NAT2*, plus rapide et plus simple d'utilisation en routine clinique. De nombreuses études ont en effet démontré une forte corrélation génotype-phénotype pour le

gène *NAT2* dans des populations de différentes origines ethniques. Notamment, le typage des sept principaux polymorphismes de ce gène est apparu suffisant pour prédire le phénotype d'acétylation des individus dans près de 100 % des cas. Le génotypage de *NAT2* apparaît donc comme un prédicteur fiable de l'activité enzymatique de NAT2 et constitue aujourd'hui la méthode de choix pour déterminer le statut d'acétylation.

Les quatre substitutions non synonymes aux positions 191, 341, 590 et 857 caractérisent les principaux clusters d'haplotypes lents de *NAT2* (respectivement *NAT2*14*, *NAT2*5*, *NAT2*6* et *NAT2*7)*. Les individus homozygotes ou hétérozygotes composites pour deux de ces haplotypes lents sont classés acétyleurs lents.

2.5 Variabilité inter-populationnelle

La prévalence du phénotype d'acétyleur lent, amplement estimée dans des populations du monde entier, est très variable selon l'origine géographique et/ou ethnique des populations. Dépassant les 82 % chez les Egyptiens, elle n'atteint pas les 10 % chez les Inuits du Canada (Evans, 1989; Kalow, 1982; Weber and Hein, 1985). Il est possible de distinguer d'un côté les populations européennes et africaines qui présentent des proportions d'acétyleurs lents et rapides à peu près équivalentes (entre 40 et 60 %), et de l'autre, les populations d'Asie du sud-est, d'Amérique et d'Océanie dans lesquelles les acétyleurs lents sont le plus souvent minoritaires (moins de 30 %). Cette différence de proportion d'acétyleurs lents s'explique par la distribution hétérogène entre les populations humaines des quatre principaux variants lents de *NAT2*. En effet, alors que le variant *NAT2*6* affiche une distribution homogène à travers le monde, les trois autres variants sont retrouvés de façon plus localisée à un niveau continental : respectivement en Europe, Asie et Amérique, et Afrique pour *NAT2*5*, *NAT2*7* et *NAT2*14* (Sabbagh et al., 2011).

Chapitre 2

Analyse des profils de différenciation génétique des populations humaines pour le gène *NAT2*

Nous l'avons vu à plusieurs reprises au cours de cette thèse, les gènes impliqués dans la réponse aux xénobiotiques sont des candidats privilégiés pour l'étude de la sélection naturelle. Le gène *NAT2* l'est tout particulièrement, de par sa participation essentielle aux mécanismes de détoxification mis en place par l'organisme pour se défendre contre les agressions chimiques. Ce gène semble avoir une histoire évolutive très complexe et différentes hypothèses de sélection, en lien avec le mode de subsistance notamment, ont été proposées pour expliquer la distribution hétérogène des variants de *NAT2* conférant le phénotype d'acétyleur lent dans les populations humaines (Luca et al., 2008; Magalon et al., 2008; Mortensen et al., 2011; Patin et al., 2006b; Sabbagh et al., 2008).

Parmi les polymorphismes conférant une faible activité enzymatique, les trois variants lents c.191G>A (rs1801279), c.341T> (rs1801280) et c.857G>A (rs1799931), qui caractérisent respectivement les clusters d'haplotypes *NAT2*14*, *NAT2*5* et *NAT2*7*, sont distribués dans les populations humaines de façon hétérogène et possèdent une valeur de F_{ST} élevée (Sabbagh et al., 2008). Nous le savons, de tels profils de différenciation génétique peuvent refléter l'action de la sélection naturelle, notamment des pressions de sélection spécifiques de populations reflétant des phénomènes d'adaptation locale. A l'inverse, le variant c.590G>A déterminant le cluster d'haplotypes *NAT2*6* (rs1799930) affiche des fréquences relativement similaires dans

diverses populations à travers le monde (Luca et al., 2008; Sabbagh et al., 2011). Ce faible niveau de différenciation génétique suggère plutôt un processus homogénéisant de sélection naturelle favorisant, par le biais de la sélection directionnelle ou de la sélection balancée, le même allèle dans différentes populations du monde. Bien que de nombreux autres variants ont été décrits dans d'autres parties du gène *NAT2* que l'unique exon codant (Mortensen et al., 2011; Patin et al., 2006b), très peu de données existent sur la distribution géographique de ces variants dans les populations humaines.

Dans ce chapitre, nous présentons l'étude des profils de différenciation génétique des populations humaines pour l'ensemble des variants du gène *NAT2* (~ 10 kb) en utilisant les données de séquence du Projet 1000 Génomes. Comme dans nos précédentes études, nous avons cherché à déterminer si les profils de différenciation génétique atypiques observés étaient dus à l'action de la sélection naturelle.

1. Résumé de l'article 3

Dans cette étude, nous avons exploré les profils de différenciation génétique de l'ensemble des variants de *NAT2* et déterminé si ces profils étaient atypiques par rapport au reste de la variation du génome par l'utilisation d'une approche *outlier*. Celle-ci se base sur les distributions empiriques pangénomiques du F_{ST} construites lors de notre précédente étude à partir des données 1000 Génomes (partie 3, chapitre 2). Neuf distributions distinctes ont été considérées en répartissant les variants des données 1000 Génomes en différentes catégories d'annotation fonctionnelle.

Les empreintes génomiques de la sélection naturelle ont été recherchées par l'utilisation de deux approches complémentaires : (1) les tests iHS et XP-EHH basés sur la structure haplotypique, qui sont adaptés à la détection des balayages sélectifs, reflets de la sélection positive récente et (2) le *D* de Tajima, test de neutralité qui permet de détecter des écarts au spectre de fréquence allélique attendu sous l'hypothèse de neutralité sélective, comme nous l'avons expliqué dans la partie introductive de cette thèse. La significativité statistique des scores obtenus pour ces trois tests de sélection a

été déterminée par comparaison avec des distributions empiriques des statistiques de tests estimées sur un jeu de 100 régions indépendantes non codantes, supposées évoluer essentiellement de façon neutre. L'effet fonctionnel de certains variants d'intérêt a été prédit grâce à la méthode *in silico* F-SNP (Lee and Shatkay, 2009).

Nos résultats n'ont pas révélé de profils de différenciation particulièrement élevés pour les variants du gène *NAT2*. En revanche, un niveau de différenciation atypiquement faible a été détecté pour cinq variants, incluant le SNP rs1799930 définissant le cluster d'haplotypes *NAT2*6* et quatre variants introniques, tous en fort déséquilibre de liaison avec lui. Ce résultat peut indiquer soit de la sélection balancée, soit de la sélection directionnelle opérant de façon similaire dans différentes populations.
Afin de rechercher si un autre gène adjacent à *NAT2* pouvait être à l'origine de ce profil atypique, nous avons étendu notre analyse à une région de 600 kb centrée sur *NAT2*. Au sein de cette région, nous avons observé que les quelques variants en déséquilibre de liaison (r^2 > 0,10) avec le variant rs1799930 qui présentaient également un score F_{ST} significativement faible étaient tous localisés soit dans une région intergénique à proximité de *NAT2*, soit dans l'intron de *NAT2*, rendant ainsi faiblement probable l'implication d'un autre gène. En outre, bien que certains variants aient une probabilité élevée d'être délétères (score FS ≥ 0,5), le score FS le plus élevé a été observé pour le SNP rs1799930, suggérant que ce variant représente la cible la plus probable du processus sélectif identifié.
De manière intéressante, des découvertes récentes ont démontré que ce variant conférait un phénotype d'acétylation très lent (Ruiz et al., 2012; Selinski et al., 2013). Ce variant a de plus été retrouvé associé à un risque augmenté de toxicité à certains médicaments ou de cancers (An et al., 2012; Huang et al., 2002; Lee et al., 2010; Leiro-Fernandez et al., 2011; Selinski et al., 2013; Teixeira et al., 2011). La corrélation importante observée entre la distribution mondiale de ce variant et le mode de subsistance des populations évoque un potentiel avantage sélectif conféré par ce variant, en réponse à des modifications des conditions environnementales, en particulier alimentaires (Sabbagh et al., 2011).

Sous cette hypothèse, il est possible que la sélection naturelle ait favorisé l'allèle A du SNP rs1799930 en différents endroits du monde au moment de l'introduction de l'agriculture et de l'élevage, par l'effet de la sélection directionnelle ou de la sélection balancée. Étant donné qu'aucun score significatif n'a été obtenu pour ce variant avec les tests iHS et XP-EHH appliqués dans chacune des 14 populations de 1000 Génomes, ce variant n'a vraisemblablement pas été soumis à un phénomène de balayage sélectif. En revanche l'application du *D* de Tajima a conduit à des scores positifs et significatifs au seuil de 5 % autour de ce variant dans 5 populations (et à des scores à la limite de la significativité dans deux supplémentaires), suggérant la présence d'un excès d'allèles à fréquence intermédiaire, compatible avec la sélection balancée. Toutefois ces résultats mis en relation avec ceux de la littérature, permettent également d'envisager que la sélection a pu agir sur le gène *NAT2* à travers un processus plus complexe de sélection directionnelle ciblant de manière simultanée différents allèles lents dans diverses populations. Ces deux scénarios sélectifs ne sont pas mutuellement exclusifs et soulignent la complexité de l'histoire évolutive du gène *NAT2*.

2. Article 3

Patillon B, Luisi P, Poloni ES, Boukouvala S, Darlu P, Génin E, Sabbagh A. *A homogenizing process of selection has maintained an 'ultra-slow' acetylation NAT2 variant in humans*. Soumis à *Human Biology*.

A homogenizing process of selection has maintained an 'ultra-slow' acetylation *NAT2* variant in humans

Running title: *NAT2**6 as a target of homogenizing selection

Patillon B[1,2,3], Luisi P[4], Poloni ES[5], Boukouvala S[6], Darlu P[7], Genin E[†,8] and Sabbagh A[*,†,1,2]

[1] IRD UMR216, Mère et enfant face aux infections tropicales, Paris, France
[2] PRES Sorbonne Paris Cité, Université Paris Descartes, Faculté de Pharmacie, Paris, France
[3] Université Paris Sud, Kremlin-Bicêtre, France
[4] Institute of Evolutionary Biology, CEXS-UPF-PRBB, Catalonia, Barcelona, Spain
[5] Laboratory of Anthropology, Genetics and Peopling History, Anthropology Unit, Department of Genetics and Evolution, University of Geneva, Geneva, Switzerland
[6] Department of Molecular Biology and Genetics, Democritus University of Thrace, Alexandroupolis, Greece
[7] CNRS UMR7206, Muséum National d'Histoire Naturelle, Université Paris Diderot, Paris, France
[8] INSERM U1078, UBO EFS Bretagne, Brest, France

*Correspondence: Dr A. Sabbagh, UMR 216 IRD - Université Paris Descartes, Faculté de Pharmacie, 4 avenue de l'Observatoire, 75270 Paris Cedex 06, France. Tel: +33 153739620; Fax: +33 153739617; E-mail: audrey.sabbagh@ird.fr
[†]These authors contributed equally to the work.

Abstract

N-acetyltransferase 2 (NAT2) is an important enzyme involved in the metabolism of a wide spectrum of naturally occurring xenobiotics, including therapeutic drugs and common environmental carcinogens. Extensive polymorphism in NAT2 gives rise to a wide interindividual variation in acetylation capacity which influences individual susceptibility to various drug-induced adverse reactions and cancers. Striking patterns of geographic differentiation have been described for the main slow acetylation variants of the *NAT2* gene, suggesting the action of natural selection at this locus. In the present study, we took advantage of the whole-genome sequence data available from the 1000 Genomes project to investigate the global patterns of population genetic differentiation at *NAT2* and determine whether they are atypical compared to the remaining variation of the genome. The non-synonymous substitution c.590G>A (rs1799930) defining the slow *NAT2*6* haplotype cluster exhibited an unusually low F_{ST} value when compared to the genome average (F_{ST} = 0.006, *P*-value = 0.016). It was pointed out as the most likely target of a homogenizing process of selection promoting the same allelic variant in globally distributed populations. The rs1799930 A allele has been associated with the slowest acetylation capacity *in vivo* and its substantial correlation with the subsistence strategy adopted by past human populations suggests that it may have conferred a selective advantage in populations shifting from foraging to agricultural and pastoral activities in the Neolithic period. Results of neutrality tests further supported an adaptive evolution of the *NAT2* gene through either balancing selection or directional selection acting on multiple standing slow-causing variants.

Keywords: *NAT2*; acetylation polymorphism; population differentiation; natural selection; linkage disequilibrium; rs1799930.

Introduction

The human acetylation polymorphism is one of the oldest and best-characterized pharmacogenetic traits that underlie interindividual and interethnic differences in response to xenobiotics. It refers to a genetically determined difference in the *N*-acetylation capacity of a variety of clinically useful drugs and known carcinogens present in the diet, cigarette smoke and the environment.[1,2] Some of the drugs excreted by acetylation are crucial in the treatment of diseases representing a worldwide concern, such as tuberculosis, AIDS-related complex diseases, and hypertension. The individual acetylation status has proven to be an important determinant of both the effectiveness of prescribed medications and the development of adverse drug reactions and toxicity during drug treatment.[3,4] Moreover, epidemiological studies have associated the acetylation phenotype with an increased susceptibility to various cancers following exposure to aromatic amine carcinogens.[5-8]

N-acetylation activity has been investigated in a wide range of populations, leading to a phenotype classification of humans in two main categories: fast acetylators, who exhibit the so-called 'wild-type' or normal acetylation activity, and slow acetylators, characterized by a decreased enzyme activity. The proportions of rapid and slow acetylators vary remarkably between populations of different ethnic and/or geographic origin.[9-11] Depending on the test substrate administered, a trimodal, rather than bimodal, distribution can be observed, revealing an additional, intermediate phenotype.[12-14] Moreover, recent results suggest that the slow acetylator phenotype is not homogeneous and that several slow acetylator phenotypes may rather exist, resulting from an allelic heterogeneity and differential functional effects of the slow acetylation alleles.[8,15] A refinement in phenotype inference, notably by the consideration of an 'ultra-slow' acetylator category, is advocated to help identifying new clinically relevant associations with one or more of these phenotype subcategories.

Acetylation polymorphism arises from allelic variations in the *NAT2* gene, which result in the production of arylamine *N*-acetyltransferase 2 (NAT2) proteins with variable enzyme activity or stability. The *NAT2* gene contains two exons with a relatively long intronic region of about 8.6 kb. Exon 1 is very short (100 bp) and the entire protein-coding region is contained within the 870-bp exon 2. Extensive polymorphism has been described in exon 2, with 38 nucleotide variations registered to date (http://nat.mbg.duth.gr/). Of these, four common non-synonymous substitutions at positions 191, 341, 590 and 857 are the most studied and characterize the major *NAT2* slow haplotype clusters (*NAT2*14*, *NAT2*5*, *NAT2*6* and *NAT2*7*, respectively). Individuals who are homozygous or compound heterozygous for two of these low-activity haplotypes are classified as slow acetylators.

Striking patterns of geographic differentiation have been described for the major *NAT2* slow-causing variants, suggesting the action of natural selection at this locus.[11,16] The function of NAT2 in mediating the interactions between humans and their chemical environment, which varies depending on diet and lifestyle, makes it an excellent candidate for population-specific selection pressures, assuming that xenobiotic exposure has significantly impacted population fitness over time.[11,17-21] Notably, an unusually high level of population differentiation between East Asians and other populations (F_{ST} values around 0.40) has been described for the c.341T>C slow-causing variant (rs1801280), as well as the two linked c.481C>T (rs1799929) and c.803A>G (rs1208) nonfunctional SNPs, when compared to an empirical distribution of F_{ST} computed across a 400-kb region encompassing the whole human *NAT* gene family.[22] In contrast, the slow 590A variant (rs1799930) was found to occur at roughly similar frequencies among widely dispersed populations.[11,19] Such a low level of geographic differentiation may rather suggest a homogenizing process of natural selection, promoting the same allelic variant in otherwise disparate populations (through either directional or balancing selection). Although many polymorphisms have been described in other regions of the *NAT2* gene,[17,21] limited data exist on the geographic distribution of these variants in worldwide populations.

In this study, we took advantage of the whole-genome sequence data available from the 1000 Genomes (1KG) project to explore the global patterns of population genetic differentiation for the whole set of variants occurring in the entire *NAT2* gene sequence (~10 kb). An outlier approach was used to determine whether the patterns of geographic differentiation at this locus were atypical compared to those observed for the remaining variation of the genome. Selection tests based on the site frequency spectrum and extended haplotype homozygosity were further applied to determine the possible role of natural selection in shaping the atypical patterns observed.

Materials and methods

Data retrieval

Whole-genome variation data generated by the 1KG project in 1,089 unrelated individuals was directly downloaded from the 1000 Genomes ftp site (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/release/20110521/), using the phase 1 integrated release version 3 of April 2012.[23] The 1,089 individuals are drawn from 14 different populations in sub-Saharan Africa, Europe, East Asia, and the Americas: Yoruba in Ibadan, Nigeria (YRI); Luhya in Webuye, Kenya (LWK); people with African ancestry in the

Southwest United States (ASW); Utah residents with Northern and Western European ancestry (CEU); Tuscans in Italy (TSI); British in England and Scotland (GBR); Finnish in Finland (FIN); Iberians in Spain (IBS); Han Chinese in Beijing, China (CHB); Southern Han Chinese in China (CHS); Japanese in Tokyo, Japan (JPT); people with Mexican ancestry in Los Angeles, California (MXL); Colombians in Medellin, Colombia (CLM); and Puerto Ricans in Puerto Rico (PUR). From the obtained vcf (variant call format) files, we extracted exclusively the low-coverage VQSR (Variant Quality Score Recalibrator method) SNV calls in order to avoid any bias that might result from differences between low-coverage whole-genome calls and high-coverage exome SNV calls. Indels were not used. Functional annotation of the 36,382,866 SNVs retrieved was performed using classification from the dbSNP database (build 137) (http://genome.ucsc.edu/cgi-bin/hgTables:SNV137.txt). SNVs were assigned to two main classes: genic and nongenic SNVs. Genic SNVs were further classified as intronic, 5'-UTR, 3'-UTR, coding synonymous, coding non-synonymous or splice-site.

Population genetic differentiation

Global levels of population genetic differentiation at *NAT2* (chr8:18248755-18258723 in the human GRCh37/hg19 assembly) were evaluated by using the fixation index F_{ST},[24] which quantifies the proportion of genetic variance explained by allele frequency differences among populations. F_{ST} ranges from 0 (for genetically identical populations) to 1 (for completely differentiated populations). F_{ST} scores were computed for all *NAT2* SNVs occurring with a minor allele frequency (MAF) $\geq$ 0.05 in at least one of the 14 1KG populations, using the BioPerl module PopGen.[25] Extreme values of F_{ST} can result from natural selection but also from nonselective events linked to the demography of populations, such as genetic drift. Because such nonselective processes randomly act on the genome, they are expected to have the same average effect across the genome, in contrast to natural selection, which impacts population differentiation in a locus-specific manner. The genome-wide variation data provided by the 1KG project can thus be used to infer the action of natural selection by adopting an outlier approach.[26] For that purpose, we built nine empirical distributions of the F_{ST} statistic by considering different subsets of SNVs defined according to their physical location and/or functional impact. To obtain distributions of likely independent observations, a LD-based pruning procedure was applied to each of these nine subsets using Plink[27] with default parameters (pruning based on a variance inflation factor of at least 2 within each sliding window of 50 SNVs with a step of five SNVs). This resulted in a total of 25,532,386 independent autosomal SNVs included in the genome-wide empirical distribution. These numbers are respectively 15,141,160, 11,282,100, 10,477,050, 24,395, 198,718, 107,644,

146,572, and 1,912 in the nongenic, genic, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice-site distributions. These nine distributions were then used as reference to assess whether the patterns of genetic differentiation observed at *NAT2* are atypical. Empirical *P*-values were estimated as the proportion of F_{ST} scores in the empirical distribution that are either higher (diversifying selection) or lower (homogenizing selection) than the value observed at the locus of interest. Since F_{ST} strongly correlates with heterozygosity,[28-30] empirical *P*-values were calculated within bins of SNVs grouped according to their global MAF. A total of 27 bins were considered for the whole MAF range: 10 bins of size 0.001 for MAF between 0 and 0.01, 9 bins of size 0.01 for MAF between 0.01 and 0.10, and 8 bins of size 0.05 for MAF between 0.10 and 0.50.

Selection tests

To determine whether natural selection has played a role in the unusual patterns of geographic differentiation disclosed, we used two complementary approaches based on the allele frequency spectrum of segregating sites and on the local haplotype structure. Tajima's *D*[31] is a classical neutrality test that compares estimates of the number of segregating sites and the average number of pairwise differences between nucleotide sequences (π). A zero value of the test statistic *D* is expected under the null hypothesis of selective neutrality, whereas a positive *D* is taken as indicative of balancing selection and a negative one of directional selection. Tajima's *D* scores were computed across the whole *NAT2* coding region by using a sliding window approach with a window size of 1 kb and a step size of 100 bp. Statistical significance of the test statistic was assessed using an empirical approach. From the genome-wide data available from the 1KG project, we selected a set of unlinked noncoding regions expected to be mostly neutrally evolving. A total of 100 autosomal regions of size 1kb were selected that met the following criteria: (i) to be at least 100 kb away from any known or predicted genes or expressed sequence tag or region transcribed into mRNA; (ii) to be outside any segmental duplication or region transcribed into a long noncoding RNA or conserved noncoding element (as defined in Woolfe *et al*[32]); (iii) to be distant from each other by at least 100 kb and not in linkage disequilibrium (LD) with each other; (iv) to contain a number of SNVs equal to the mean number of SNVs included in the 1-kb sliding windows spanning the *NAT2* coding region. Tajima's *D* scores were computed for these 100 regions so as to obtain the null (neutral) distribution of the test statistic in each population sample. An empirical *P*-value was estimated at each sliding window position within *NAT2* by considering the proportion of regions showing a test statistic greater (excess of intermediate-frequency variants) or lower (excess of low-frequency variants) than the value observed at that specific position.

We next used methods based on the extended haplotype homozygosity (EHH) measure, *i.e.* the sharing of identical alleles across relatively long distances by most haplotypes in a population sample.[33] We calculated the integrated haplotype score (iHS)[34] that compares the rate of EHH decay observed for both the derived and ancestral allele at each core SNV. An extremely positive or negative value at the core SNV provides evidence of positive selection with unusually long haplotypes carrying the ancestral or the derived allele, respectively. The raw iHS scores were computed for all *NAT2* SNVs using the iHS option implemented in the WHAMM! software,[34] which we slightly modified in order to speed up computation times: thresholds for EHH decay were modified from 0.25 to 0.15 and the size of the analyzed region was set to 0.2 Mb instead of 2.5 Mb. Information on ancestral allele state was obtained from a four-way alignment of human, chimpanzee, orangutan and rhesus macaque species, provided by the 1KG consortium (ftp://ftp.1000genomes.ebi.ac.uk/vol1/ftp/phase1/analysis_results/supporting/ancestral_alignments/).

We also applied a cross-population test by computing the XP-EHH statistic[35] that compares the integrated EHH computed in a test population versus that of a reference population. XP-EHH scores were computed using the same EHH decay parameters and window size as for iHS. The Yoruba (YRI) sample was used as a reference for samples outside Africa, and the Utah residents of European ancestry (CEU) as a reference for African samples.

iHS and XP-EHH scores were also computed for all available SNVs in the 100 neutral regions described above, thus providing a reference distribution for each test statistic to estimate empirical *P*-values. Raw scores of iHS and XP-EHH in *NAT2* were standardized in bins of derived allele frequency (step size of 0.05) using the corresponding distribution of each statistic. For both iHS and XP-EHH, we used the phased data provided by the integrated Phase 1 release version 3 of the 1KG project (April 2012) and genetic distances were obtained from the high density genetic combined map based on 1KG pilot 1 data.

In-silico prediction of SNV's functional effects

The F-SNP method[36] (http://compbio.cs.queensu.ca/F-SNP/) was applied to assess the potential functional effect of SNVs. This integrative scoring method combines assessments from 16 independent computational tools and databases, using a probabilistic framework that takes into account both the certainty of each prediction and the reliability of the different tools depending on the physical and functional annotation of the specific variant tested. It provides a functional significance (FS) score that quantitatively measures the possible deleterious effect of the tested SNV at the splicing, transcriptional, translational and post-translational levels. A FS score of 0.5 is considered as the cutoff point for predicting a deleterious effect.[37]

Results and discussion

Global patterns of population genetic differentiation were examined in the genomic region spanning the entire *NAT2* gene (~10 kb), using the F_{ST} statistic[24] and the sequence variation data provided by the 1KG project.[23] *P*-values were estimated from empirical distributions built from the background genomic variation (see Materials and Methods). We assigned to each genetic variant of the *NAT2* gene a 'main *P*-value' derived from the genome-wide empirical distribution and a 'subset *P*-value' derived from the distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, nongenic, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice site).

No SNVs in *NAT2* exhibited significantly high F_{ST} values compared to the genomic background, when considering the global differentiation among the 14 worldwide populations from 1KG (all *P*-values > 0.05). No significant *P*-values were observed when contrasting East Asia to the rest of the world either (data not shown). Although high F_{ST} scores were observed for the three SNVs c.341T>C (rs1801280), c.481C>T (rs1799929) and c.803A>G (rs1208) in this specific pairwise comparison ($F_{ST} \approx 0.30$), they could not be considered as atypical when compared to the rest of the genome (*P*-values ranging from 0.06 to 0.09). This contrasts with previous findings pointing out an atypical pattern of differentiation for these three variants when considering HapMap data and only the set of variants located within a 400-kb region surrounding the *NAT2* gene as a reference distribution.[22] This difference may be due to the different set of populations surveyed or to the more accurate empirical distribution used to represent the background genomic variation in the present study.

In contrast, five *NAT2* SNVs exhibited unusually low F_{ST} values when compared to the genome average, with both main and subset *P*-values below 0.05 (Figure 1). Four of them (rs6984200, rs2087852, rs11996129 and rs1112005) are located in the intronic region of *NAT2*, while the fifth one (rs1799930) is a non-synonymous substitution defining the *NAT2*6* slow haplotype cluster (c.590G>A resulting in R197Q). Note that the four intronic SNVs are in high LD with the rs1799930 variant (r^2 ranging from 0.80 to 0.88) (Table 1). They all occur at high frequencies in the global human population (within the 0.23-0.26 MAF range). Such low levels of population genetic differentiation suggest that at least one of these polymorphisms may be subject to balancing or species-wide directional selection, the rs1799930 being the most likely target given its gene location and functional impact.

To determine whether another putative candidate in the genomic region surrounding *NAT2* might explain the patterns observed, through a significant LD with these variants, we extended the analysis to a 600-kb region centered on the human *NAT* multigene family on

chromosome 8 (Figure 2). All the variants exhibiting an r^2 value above the 0.10 threshold with the rs1799930 SNV are located within a 56-kb region (chr8:18229877-18285763 in hg19) in the direct vicinity of the *NAT2* gene (Figure 2A), making unlikely the involvement of another gene in the region. Table 1 provides the list of variants showing a significantly low interpopulation F_{ST} value for either the main or subset *P*-value and being in moderate to strong LD with the rs1799930 variant ($r^2 > 0.50$). They are all either intergenic (located up to ~3 kb upstream or 22 kb downstream of the *NAT2* gene) or intronic to *NAT2*. The prediction of their functional impact with the F-SNP method revealed high FS scores for some of them (FS = 0.50), denoting a potentially deleterious effect by affecting either the transcriptional or splicing regulation. However, the highest FS score was observed for the rs1799930 polymorphism (0.87), which also displayed the lowest subset *P*-value ($P = 0.016$; Table 1). Altogether, these results point to the rs1799930 polymorphism as the most likely target of homogenizing selection in the genomic region surveyed.

Interestingly, recent evidence suggest that the *NAT2*6* haplotype cluster (characterized by the rs1799930 A slow-causing allele) is related with the slowest acetylation capacity *in vivo*, and that the homozygous genotype *NAT2*6/*6* thus defines a new category of 'ultra-slow' acetylators.[8,15] Ultra-slow acetylators have about 30% lower activities of caffeine metabolism compared with other slow acetylators. This is of the same order of magnitude than the reduction in enzyme activity between rapid and intermediate acetylators.[15] These findings are consistent with a previous study by Cascorbi *et al*[13] that demonstrated a markedly decreased NAT2 activity *in vivo* in *NAT2*6/*6* compared to *NAT2*5/*5* genotypes. Indirect evidence is also provided by clinical association studies related to both drug toxicity and cancer risk. Anti-tuberculosis drug-induced hepatotoxicity risk has been shown to be particularly high in carriers of the *NAT2*6/*6* genotype.[38-43] Similarly, the ultra-slow genotype, and not the common slow *NAT2* genotype, has been significantly associated with an increased risk of urinary bladder cancer.[8] This could explain why this particular *NAT2* slow-causing variant, and not another one, may have been a specific target of natural selection.

Although a highly homogenous distribution of the rs1799930 A allele is observed across worldwide populations (with a global frequency of 0.246), resulting in an unusually low F_{ST} value (interpopulation $F_{ST} = 0.006$), a three-fold lower frequency of this allele has been reported in hunter-gatherers (~0.08) as compared to agriculturalists and pastoralists (~0.25) in a comprehensive survey of human *NAT2* variation including 128 population samples classified according to their major subsistence strategy.[11] This significant difference in the frequency of *NAT2*6* alleles ($P < 0.0001$) was identified as the main genetic cause of the higher prevalence of the slow acetylation phenotype in populations practicing farming and herding as compared to those mostly relying on hunting and gathering (46% vs 22%,

respectively).[11] Given this marked correlation between the rs1799930 A allele and the subsistence strategy adopted by past populations in the last 10,000 years, it has been suggested that this slow-causing allele may have conferred a selective advantage in populations shifting from foraging to agricultural and pastoral activities in the Neolithic period. New or more concentrated NAT2 substrates introduced in the chemical environment of food-producing communities have likely promoted a slower acetylation rate in these populations. Consequently, the markedly low level of population differentiation observed at the rs1799930 locus may result from the convergent selection of the rs1799930 A allele in agriculturalist and pastoralist populations which are now present in most parts of the world.

Several lines of evidence support the hypothesis that the rs1799930 G>A non-synonymous substitution (R197Q) has specifically occurred in the human lineage. First, the NAT2 197R residue appears to be highly conserved throughout primate evolution, with 100% of the orthologous *NAT2* sequences generated in 19 distinct simian species harboring an arginine (R) at this position.[44] Second, the 197R position was found to be monomorphic in 103 individuals from six great ape species (*Pan troglodytes*, *Pan paniscus*, *Gorilla beringei*, *Gorilla gorilla*, *Pongo abellii*, *Pongo pygmaeus*) (Prado-Martinez *et al*,[45] E.S. Poloni, personal communication), as well as in 28 Rhesus monkeys (*Macaca mulatta*) fully sequenced for the *NAT2* gene (A. Sabbagh, personal communication), making the R197Q polymorphism a specific feature of the human lineage. The hypothesis of a trans-species polymorphism maintained for several million years, through shared balancing selection pressures, seems therefore unlikely.

Assuming that the rs1799930 A allele has conferred a selective advantage to populations shifting from food collection to farming and animal breeding in the Holocene, this could have happened either through directional selection or balancing selection. A gene-dose effect has been indeed described for this variant, with a significant trend toward a slower acetylation capacity in individuals carrying an increasing number of *NAT2*6* haplotypes (0, 1 or 2).[8,15] Therefore, heterozygous individuals for this allele display an intermediate metabolic phenotype that may have been advantageous if one considers the competing needs of both maintaining an efficient detoxification of harmful xenobiotics and avoiding the damaging effects of the putative carcinogens that can be activated through NAT2 acetylation. In an attempt to provide further insights into the evolutionary mechanisms that might have driven and maintained the rs1799930 A allele at high frequencies in most human populations worldwide, we carried out several tests of natural selection based on the allele frequency spectrum (Tajima's D[31]) and haplotype structure (iHS[34] and XP-EHH[35]). An empirical approach using sequence variation data from 100 unlinked noncoding regions was adopted to assess statistical significance.

All iHS and XP-EHH scores computed for the rs1799930 SNV in all individual populations from 1KG were not significant at the 0.05 threshold (Table 2). This precludes a clear signal of positive selection for this variant as the one expected under a 'hard sweep model', which assumes the rapid fixation of a single newly arisen advantageous mutation.[46] In contrast, we found significant Tajima's *D* scores in the 1-kb regions encompassing the rs1799930 variant in five population samples: British, Finnish, Tuscans, Utah residents of European ancestry and Puerto Ricans ($P < 0.05$) (Table 2, Supplementary Table 1). We acknowledge that these results become non-significant when a correction for multiple testing is applied, but we also note that the ratio of five significant tests out of 14 is higher than the expected 5% proportion of false positives. Non-significant scores, but with *P*-values getting closer to the 5% threshold, were observed in two additional ones ($P = 0.07$ and $P = 0.08$ in Colombians and Luhya, respectively). Furthermore, although non-significant scores prevent rejection of the null hypothesis of selective neutrality, it is noteworthy that all populations tested but one (Japanese) gave positive Tajima's *D* values, suggesting a trend toward an excess of intermediate-frequency variants compatible with the action of balancing selection. Such consistent results for populations with different demographic pasts make it unlikely that they are due to demography rather than balancing selection. This is also in agreement with previous findings demonstrating globally positive and significant Tajima's *D* values in different continental populations and the absence of any signature of positive selection, as detected by EHH-based tests.[19-21,47] A notable exception concerns the c.341C>T slow-causing variant for which a selective sweep was detected in Western and Central Eurasian populations[17] with the long-range haplotype test.[33] We did not confirm such signature of positive selection at this locus in any of the 14 populations from 1KG (both iHS and XP-EHH scores not significant at the 0.05 level). No significant scores were observed for any of the other slow-causing variants either (c.191G>A, c.341T>C and c.857G>A; data not shown). Therefore, patterns of diversity at *NAT2* seem compatible with either balancing selection or a more complex model of 'multiallelic' directional selection where different slow variants of *NAT2* may have simultaneously become targets of directional selection, thereby generating an excess of intermediate-frequency alleles. This would explain why our conventional tests of selection based on EHH, more suited to detect classical selective sweeps, failed to detect a signature of positive selection at the rs1799930 locus. The signature of selection at this individual position could have been weakened by the global increase in frequency of other *NAT2* altering mutations. Note, however, that contrary to c.191G>A (*NAT2*14*), c.341T>C (*NAT2*5*) and c.857G>A (*NAT2*7*), which mainly cluster in specific continental regions (sub-Saharan Africa, Europe and Asia, respectively,[11] the cosmopolitan distribution of the c.590G>A variant (*NAT2*6*) suggests that it may have been positively selected in globally

distributed food-producing communities. Finally, the hypotheses of balancing and directional selection are not mutually exclusive and multiple modes of selection may have operated at the *NAT2* locus on a population-specific basis, as previously suggested.[21]

In conclusion, we have described an atypical pattern of geographic differentiation for five genetic variants of the *NAT2* gene, including the functional rs1799930 SNP defining the slow *NAT2*6* haplotype series, and four intronic SNPs in high LD with it. An extended analysis of a 600-kb region surrounding *NAT2* pointed to the rs1799930 polymorphism as the most likely target of a homogenizing process of natural selection promoting the same allelic variant in most human populations, resulting in an unusually low F_{ST} value (F_{ST} = 0.006). The rs1799930 A allele has been associated with the slowest acetylation capacity *in vivo* and is much more frequent in agriculturalists and pastoralists as compared to hunter-gatherers, suggesting it may have been positively selected in food-producing communities which are now present in most parts of the world. Neutrality tests based on the allele frequency spectrum revealed a trend toward an excess of intermediate-frequency variants at *NAT2*, compatible with either balancing selection or a more complex model of multiallelic directional selection. Our findings provide further insights into the functional importance of the rs1799930 polymorphism and the role it may have played in human adaptation to fluctuating xenobiotic environments.

References

1 Butcher NJ, Boukouvala S, Sim E, Minchin RF: Pharmacogenetics of the arylamine Nacetyltransferases. Pharmacogenomics J 2002; 2: 30-42.

2 Hein DW: Molecular genetics and function of NAT1 and NAT2: role in aromatic amine metabolism and carcinogenesis. Mutat Res 2002; 506: 65-77.

3 Meisel P: Arylamine N-acetyltransferases and drug response. Pharmacogenomics 2002; 3: 349–366.

4 Ladero JM: Influence of polymorphic N-acetyltransferases on non-malignant spontaneous disorders and on response to drugs. Curr Drug Metab 2008; 9: 532-537.

5 Hein DW: N-acetyltransferase 2 genetic polymorphism: effects of carcinogen and haplotype on urinary bladder cancer risk. Oncogene 2006; 25: 1649-1658.

6 Sanderson S, Salanti G, Higgins J: Joint effects of the N-acetyltransferase 1 and 2 (NAT1 and NAT2) genes and smoking on bladder carcinogenesis: a literature-based systematic HuGE review and evidence synthesis. Am J Epidemiol 2007; 166: 741-751.

7 Agúndez JA: Polymorphisms of human N-acetyltransferases and cancer risk. Curr Drug Metab 2008; 9: 520-531.

8 Selinski S, Blaszkewicz M, Ickstadt K, Hengstler JG, Golka K: Refinement of the prediction of N-acetyltransferase 2 (NAT2) phenotypes with respect to enzyme activity and urinary bladder cancer risk. Arch Toxicol 2013; 87: 2129-2139.

9 Weber WW, Hein DW: N-acetylation pharmacogenetics. Pharmacol Rev 1985; 37: 25- 79.

10 Walker K, Ginsberg G, Hattis D, Johns DO, Guyton KZ, Sonawane B: Genetic polymorphism in N-Acetyltransferase (NAT): Population distribution of NAT1 and NAT2 activity. J Toxicol Environ Health B Crit Rev 2009; 12: 440-472.

11 Sabbagh A, Darlu P, Crouau-Roy B, Poloni ES: Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. PLoS One 2011; 6: e18507.

12 Kilbane AJ, Silbart LK, Manis M, Beitins IZ, Weber WW: Human N-acetylation genotype determination with urinary caffeine metabolites. Clin Pharmacol Ther 1990; 47: 470–477.

13 Cascorbi I, Drakoulis N, Brockmoller J, Maurer A, Sperling K, Roots I: Arylamine Nacetyltransferase (NAT2) mutations and their allelic linkage in unrelated Caucasian individuals: correlation with phenotypic activity. Am J Hum Genet 1995; 57: 581–592.

14 Parkin DP, Vandenplas S, Botha FJ et al: Trimodality of isoniazid elimination: phenotype and genotype in patients with tuberculosis. Am J Respir Crit Care Med 1997; 155: 1717– 1722.

15 Ruiz JD, Martínez C, Anderson K et al: The differential effect of NAT2 variant alleles permits refinement in phenotype inference and identifies a very slow acetylation genotype. PLoS One 2012; 7: e44629.

16 García-Martín E: Interethnic and intraethnic variability of NAT2 single nucleotide polymorphisms. Curr Drug Metab 2008; 9: 487-497.

17 Patin E, Barreiro LB, Sabeti PC et al: Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. Am J Hum Genet 2006; 78: 423– 436.

18 Sabbagh A, Darlu P: SNP selection at the NAT2 locus for an accurate prediction of the acetylation phenotype. Genet Med 2006; 8: 76-85.

19 Luca F, Bubba G, Basile M et al: Multiple advantageous amino acid variants in the NAT2 gene in human populations. PLoS One 2008; 3: e3136.
20 Magalon H, Patin E, Austerlitz F et al: Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. Eur J Hum Genet 2008; 16: 243–251.
21 Mortensen HM, Froment A, Lema G et al: Characterization of genetic variation and natural selection at the arylamine N-acetyltransferase genes in global human populations. Pharmacogenomics 2011; 12: 1545-1558.
22 Sabbagh A, Langaney A, Darlu P, Gérard N, Krishnamoorthy R, Poloni ES: Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet 2008; 9: 21.
23 The 1000 Genomes Project Consortium, Abecasis GR, Auton A et al: An integrated map of genetic variation from 1,092 human genomes. Nature 2012; 491: 56–65.
24 Wright S: The genetical structure of populations. Ann Eug 1951; 15: 323–354.
25 Stajich JE, Block D, Boulez K et al: The Bioperl toolkit: Perl modules for the life sciences. Genome Res 2002; 12: 1611–1618.
26 Kelley JL, Madeoy J, Calhoun JC, Swanson W, Akey JM: Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res 2006; 16: 980–989.
27 Purcell S, Neale B, Todd-Brown K et al: PLINK: a toolset for whole-genome association and population-based linkage analysis. Am J Hum Genet 2007; 81: 559–575.
28 Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L: Natural selection has driven population differentiation in modern humans. Nat Genet 2008; 40: 340–345.
29 Beaumont MA, Nichols RA: Evaluating loci for use in the genetic analysis of population structure. Proc R Soc Lond B 1996; 263: 1619–1626.
30 Elhaik E: Empirical distributions of FST from large-scale human polymorphism data. PLoS One 2012; 7: e49837.
31 Tajima F: Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 1989; 123: 585–595.
32 Woolfe A, Goode DK, Cooke J et al: CONDOR: a database resource of developmentally associated conserved non-coding elements. BMC Dev Biol 2007; 7: 100.
33 Sabeti PC, Reich DE, Higgins JM et al: Detecting recent positive selection in the human genome from haplotype structure. Nature 2002; 419: 832–837.
34 Voight BF, Kudaravalli S, Wen X, Pritchard JK: A map of recent positive selection in the human genome. PLoS Biol 2006; 4: e72.
35 Sabeti PC, Varilly P, Fry B et al: Genome-wide detection and characterization of positive selection in human populations. Nature 2007; 449: 913–918.
36 Lee PH, Shatkay H: F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res 2008; 36: D820–D824.
37 Lee PH, Shatkay H: An integrative scoring system for ranking SNPs by their potential deleterious effects. Bioinformatics 2009; 25: 1048–1055.
38 Huang YS, Chern HD, Su WJ et al: Polymorphism of the N-acetyltransferase 2 gene as a susceptibility risk factor for antituberculosis drug-induced hepatitis. Hepatology 2002; 35: 883–889.
39 Lee SW, Chung LS, Huang HH, Chuang TY, Liou YH, Wu LS: NAT2 and CYP2E1 polymorphisms and susceptibility to first-line anti-tuberculosis drug-induced hepatitis. Int J Tuberc Lung Dis 2010; 14: 622–626.

40 Leiro-Fernandez V, Valverde D, Vazquez-Gallardo R et al: N-acetyltransferase 2 polymorphisms and risk of anti-tuberculosis drug-induced hepatotoxicity in Caucasians. Int J Tuberc Lung Dis 2011; 15: 1403–1408.

Conflict of interest

The authors declare no conflict of interest.

Acknowledgements

This work was financially supported by the Institut Medicament-Toxicologie-Chimie. Environnement (IMTCE). BP is supported by a PhD fellowship from the doctoral program in Public Health from Paris Sud University. PL is supported by a PhD fellowship from 'Acción Estratrégica de Salud, en el Marco del Plan Nacional de Investigacio´n Científica, Desarrollo e Innovación Tecnológica 2008–2011' from Instituto de Salud Carlos III.

Table 1. List of variants in the 600-kb region surrounding the *NAT2* gene with significantly low F_{ST} values and in high linkage disequilibrium with rs1799930

SNV	Physical position on chr 8 (hg19)	Functional annotation	Distance from *NAT2* (kb)	Global MAF	Inter-population F_{ST}	Main *P*-value	Subset *P*-value	r^2 with rs1799930	Functional significance score (F-SNP)
rs11992530	18246133	Intergenic	2.6	0.254	0.006	0.024	0.024	0.80	no known function
rs4646241	18246696	Intergenic	2.1	0.247	0.007	0.032	0.034	0.83	no known function
rs4646242	18247449	Intergenic	1.3	0.246	0.006	0.025	0.024	0.84	0.50
rs4646244	18247718	Intergenic	1.0	0.246	0.006	0.026	0.025	0.84	no known function
rs6984200	18250317	Intronic	0	0.255	0.007	0.028	0.028	0.80	0.50
rs2087852	18251926	Intronic	0	0.235	0.007	0.032	0.031	0.86	0.50
rs11996129	18254575	Intronic	0	0.236	0.005	0.021	0.020	0.85	0.37
rs1112005	18255876	Intronic	0	0.227	0.005	0.019	0.019	0.88	0.50
rs1799930	18258103	Coding non-synonymous	0	0.246	0.006	0.026	0.016	-	0.87
rs4646247	18258908	Intergenic	0.2	0.247	0.006	0.025	0.024	0.99	0.50
rs721398	18259305	Intergenic	0.6	0.244	0.005	0.019	0.020	0.97	no known function
rs45605031	18260239	Intergenic	1.5	0.248	0.006	0.024	0.026	0.98	no known function
rs872233	18280686	Intergenic	22.0	0.274	0.007	0.030	0.032	0.76	no known function

SNV, single nucleotide variant; MAF, minor allele frequency. Only variants with $r^2 > 0.50$ with rs1799930 are shown. SNVs located in the *NAT2* gene are shaded in grey. A functional significance score of 0.5 is considered as the cutoff point for predicting a deleterious effect.[37]

Table 2. Results of selection tests for the *NAT2* rs1799930 polymorphism

	Africa			Europe					Asia			America		
	YRI	LWK	ASW	IBS	TSI	CEU	GBR	FIN	JPT	CHB	CHS	CLM	MXL	PUR
iHS	0.18	-0.27	-0.05	-0.82	-0.08	0.17	0.02	-0.19	0.18	-0.11	0.06	0.33	0.09	-0.29
XP-EHH	-0.80	0.11	-0.50	1.38	0.54	0.80	0.68	0.42	0.20	0.42	0.02	0.19	0.07	0.13
Tajima's *D* [a]	0.33	1.00*	0.60	1.07	**2.58****	**2.49****	**3.28*****	**2.67****	-0.03	0.91	0.88	1.65*	0.99	**1.55****

Significant scores at the 0.05 threshold are shown in bold. YRI: Yoruba from Ibadan, Nigeria; LWK: Luhya fromWebuye, Kenya; ASW: people of African ancestry from the southwestern United States; IBS: Iberian populations from Spain; TSI: Tuscans from Italy; CEU: Utah residents with Northern and Westhern European ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; JPT: Japanese from Tokyo, Japan; CHB: Han Chinese from Beijing; CHS: Han Chinese from South China; CLM: Colombians from Medellín, Colombia; MXL: people of Mexican ancestry from Los Angeles, California; PUR: Puerto Ricans from Puerto Rico.

* $P \leq 0.10$, ** $P \leq 0.05$, *** $P \leq 0.01$

[a]Highest Tajima's *D* score observed in the 1-kb sliding windows spanning the rs1799930 nucleotide position in each individual population. Note that, except in Asians (Japanese and both Han Chinese samples), Luhya and African-Americans, these scores also correspond to the highest values observed across the whole *NAT2* coding region (Supplementary Table 1).

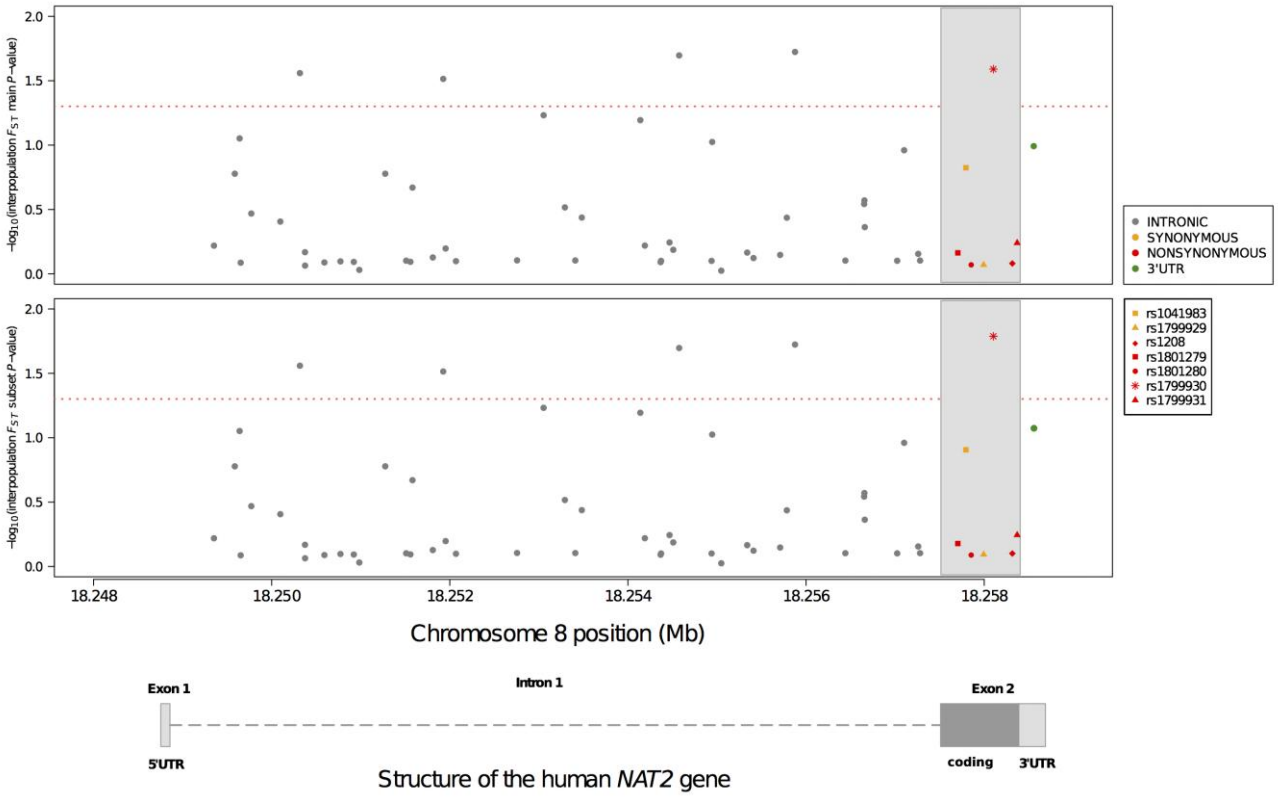


Figure 1. Distribution of -log$_{10}$ (*P*-values) of interpopulation F_{ST} scores across the human *NAT2* gene. F_{ST} scores were computed among the 14 worldwide populations from the 1000 Genomes project. *P*-values were estimated from the lower tail of the empirical distributions. (A) Main *P*-value derived from the genome-wide empirical distribution (including 25,532,386 SNVs). (B) Subset *P*-value derived from the empirical distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, intronic, 3'UTR, coding synonymous, and coding non-synonymous). The red dotted line indicates the 0.05 significance threshold.

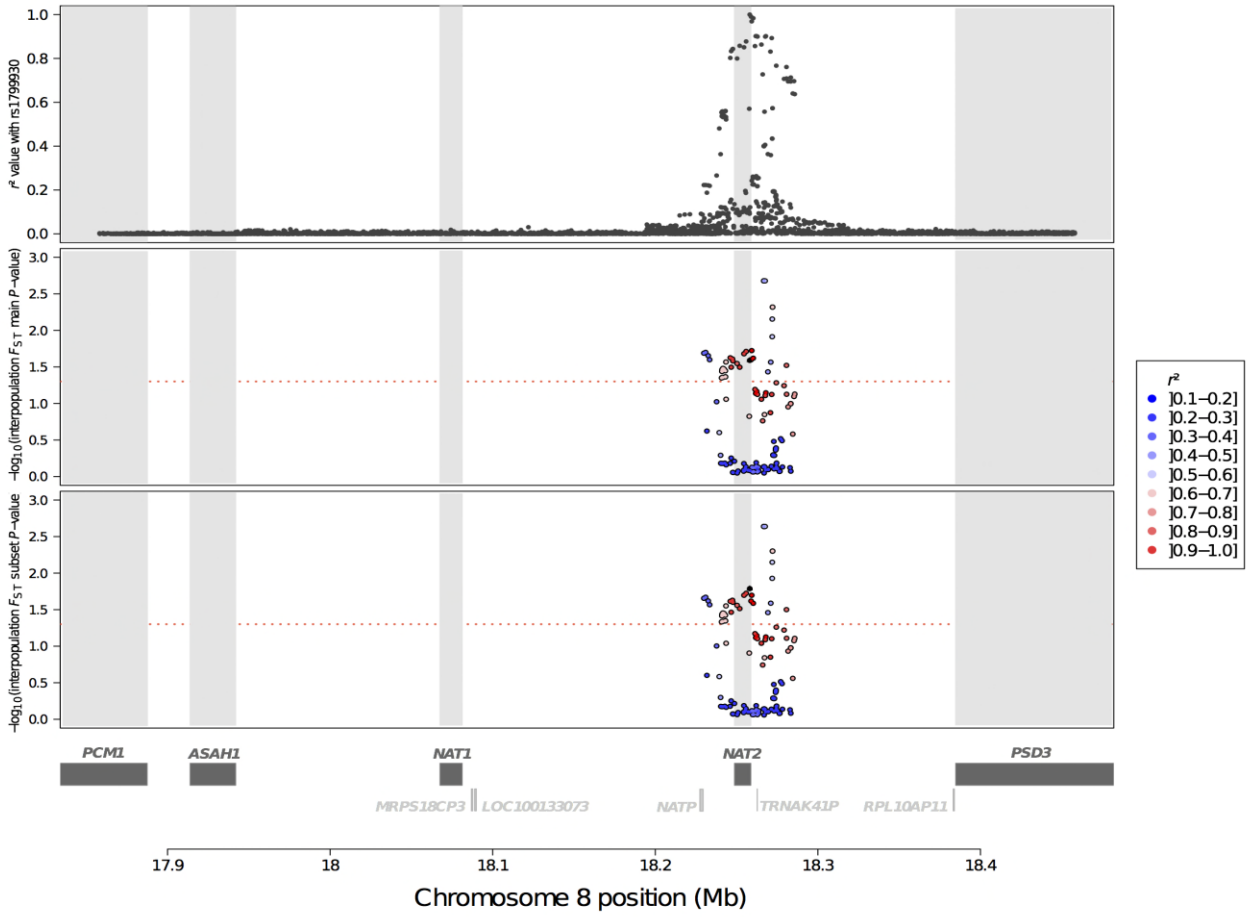


Figure 2. Distribution of -log$_{10}$ (*P*-values) of interpopulation F_{ST} scores across a 600-kb region centered on the human *NAT* gene family for those variants in linkage disequilibrium (r^2 > 0.10) with the rs1799930 polymorphism. (A) Level of linkage disequilibrium, as measured by the r^2 statistic,[48] with the rs1799930 genetic variant. **(B) Main *P*-value** estimated from the lower tail of the genome-wide empirical distribution (including 25,532,386 SNVs). **(C) Subset *P*-value** estimated from the lower tail of the empirical distribution including the subset of SNVs having a similar location and/or functional impact than the SNV of interest (*i.e.*, nongenic, intronic, 5'UTR, 3'UTR, coding synonymous, coding non-synonymous and splice site). The red dotted line indicates the 0.05 significance threshold. SNVs are displayed in different colors according to their r^2 value with rs1799930, ranging from dark blue (r^2 = 0.10) to dark red (r^2 = 1.0). The rs1799930 polymorphism is represented as a black triangle. Coding genes and pseudogenes in the 600-kb region are represented below as dark grey and light grey boxes, respectively. The genomic position (in megabases) on chromosome 8 is indicated on the horizontal axis (human GRCh37/hg19 assembly)

Supplementary Table 1. Tajima's *D* scores in the 1-kb sliding windows spanning the whole *NAT2* coding region. YRI: Yoruba from Ibadan, Nigeria; LWK: Luhya fromWebuye, Kenya; ASW: people of African ancestry from the southwestern United States; IBS: Iberian populations from Spain; TSI: Tuscans from Italy; CEU: Utah residents with Northern and Westhern European ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; JPT: Japanese from Tokyo, Japan; CHB: Han Chinese from Beijing; CHS: Han Chinese from South China; CLM: Colombians from Medellín, Colombia; MXL: people of Mexican ancestry from Los Angeles, California; PUR: Puerto Ricans from Puerto Rico.

Start[a]	End[a]	Slow-causing variant(s) included in the window[b]	Africa			Europe					Asia			America		
			YRI	LWK	ASW	IBS	TSI	CEU	GBR	FIN	JPT	CHB	CHS	CLM	MXL	PUR
18256555	18257555	none	0.12	-0.37	-0.40	-0.10	0.38	0.21	0.52	-0.45	-0.96	-1.17	-1.41	-0.27	-0.91	-0.23
18256655	18257655	none	0.16	-0.31	-0.32	-0.10	0.38	0.58	0.88	-0.26	-0.96	-1.17	-1.41	0.03	-0.78	0.32
18256755	18257755	c.191G>A	0.55	-0.28	0.04	0.46	0.38	0.78	0.92	0.02	-1.32	-1.05	-1.09	0.08	-0.67	0.23
18256855	18257855	c.191G>A, c.341T>C	1.22*	0.55	0.84	0.67	1.32	1.71	1.85	0.96	-0.61	-0.24	-0.26	0.93	0.09	1.10
18256955	18257955	c.191G>A, c.341T>C	0.94	0.41	0.55	0.67	1.32	1.71	1.85	0.96	-0.61	-0.24	-0.26	0.93	0.09	0.73
18257055	18258055	c.191G>A, c.341T>C	0.69	0.36	0.43	0.67	1.32	1.72	2.46**	1.35	-0.86	-0.24	-0.26	0.91	0.09	0.71
18257155	18258155	c.191G>A, c.341T>C, c.590G>A	0.33	0.82	0.60	0.58	1.57	1.95	3.28***	2.01*	-0.42	0.11	0.18	1.50*	0.13	1.14
18257255	18258255	c.191G>A, c.341T>C, c.590G>A	0.29	0.83	0.60	0.58	2.02	2.49**	3.28***	2.01*	-0.42	0.11	0.18	1.50*	0.13	1.14
18257355	18258355	c.191G>A, c.341T>C, c.590G>A	0.31	0.73	0.29	0.58	2.58**	2.49**	3.28***	2.01*	-0.42	0.11	0.19	1.50*	0.67	1.55**
18257455	18258455	c.191G>A, c.341T>C, c.590G>A, c.857G>A	0.17	1.00*	0.19	1.07	2.06	1.95	2.64***	2.67**	-0.37	0.35	0.78	1.65*	0.96	1.37*
18257555	18258555	c.191G>A, c.341T>C, c.590G>A, c.857G>A	-0.007	0.53	0.008	0.70	1.71	1.61	2.23*	2.18*	-0.12	0.35	0.78	1.28	0.99	1.13
18257655	18258655	c.191G>A, c.341T>C, c.590G>A, c.857G>A	-0.007	0.53	0.008	0.70	1.71	1.61	2.23*	2.18*	-0.12	0.35	0.78	1.28	0.99	1.13
18257755	18258755	c.341T>C, c.590G>A, c.857G>A	-0.04	0.58	0.06	0.70	2.16*	1.61	2.23*	2.18*	-0.12	0.35	0.41	1.28	0.99	1.51*
18257855	18258855	c.590G>A, c.857G>A	-0.70	-0.18	-0.62	0.54	0.96	0.87	1.51	1.44	-0.49	0.01	-0.007	0.61	0.03	0.83
18257955	18258955	c.590G>A, c.857G>A	-0.42	0.05	-0.47	0.45	1.27	1.21	1.81	1.73	-0.03	0.32	0.39	0.83	0.36	0.97
18258055	18259055	c.590G>A, c.857G>A	-0.82	-0.78	-0.88	0.10	0.22	-0.11	0.68	0.62	0.64	0.91	0.88	0.05	-0.22	0.21
18258155	18259155	c.857G>A	-0.67	-1.12	-1.05	0.14	-0.17	-0.51	0.29	0.23	0.23	0.70	0.54	-0.23	-0.03	0.004
18258255	18259255	c.857G>A	-0.47	-1.01	-0.92	0.14	-0.17	-0.51	0.29	0.23	-0.16	0.70	0.54	-0.23	-0.03	0.004
18258355	18259355	c.857G>A	-0.54	-0.90	-0.77	-0.79	-0.35	-0.61	0.12	0.07	0.58	1.42	1.20	-0.13	-0.51	-0.27

$^{*}P \leq 0.10$, $^{**}P \leq 0.05$, $^{***}P \leq 0.01$

[a]Genomic position of each 1-kb window on chromosome 8 in the human GRCh37/hg19 assembly.

[b]The four slow causing variants c.191G>A, c.341T>C, c.590G>A and c.857G>A correspond to the rs1801279, rs1801280, rs1799930, rs1799931 SNPs, respectively.

Discussion

Comme l'a fait remarquer l'un des pères fondateurs de la pharmacogénétique, Werner Kalow, à la fin de sa carrière, la pharmacogénétique ne permettra jamais une prédiction parfaite de la réponse d'un individu aux thérapies médicamenteuses. L'effet du médicament sur l'individu n'est jamais exactement le même, celui de l'organisme sur le médicament non plus. L'une des principales raisons à cela est que l'expression des différents gènes impliqués dans la réponse aux médicaments n'est pas constante. Entrent en jeu dans cette variabilité des facteurs à la fois internes (alimentation, maladie, variations hormonales, alcool, ...) et environnementaux (mécanismes épigénétiques, interactions gène-environnement, ...). En conséquence, un effet *in vivo* identique chez tous les individus porteurs d'une même mutation génétique ne pourra jamais être totalement garanti et il y aura toujours une part de la variabilité de réponse aux médicaments qui échappera à notre connaissance et à notre capacité d'anticipation.

Par ailleurs, l'utilité de la pharmacogénétique peut sembler discutable pour certaines molécules potentiellement dangereuses, comme les AVK. On pourrait penser en effet qu'il serait plus opportun, plutôt que de dépenser du temps et de l'argent à mettre au point un test pharmacogénétique permettant de rendre plus sûre l'utilisation de ces molécules, de développer de nouveaux médicaments, mieux tolérés et plus faciles à utiliser et présentant la même efficacité. Ainsi, une nouvelle génération d'anticoagulants oraux a été développée et introduite en France depuis 2008 : les « -xabans » (rivaroxaban (Xarelto®), apixaban (Eliquis®)), qui inhibent la voie commune de la cascade de la coagulation en neutralisant directement le facteur X activé (Xa), et les « -gatrans » (dabigatran (Pradaxa ®)), qui agissent à la fin de cette voie en inhibant directement le facteur II activé (IIa). Opérant ainsi par un mécanisme pharmacologique différent des AVK, ils promettent le même bénéfice clinique que ces derniers avec une marge thérapeutique plus large qui permet d'éviter un suivi biologique aussi lourd. Si ces molécules tiennent leurs promesses, l'usage d'un test génétique pour ces molécules ne sera pas justifié et toutes les connaissances acquises ces dernières années sur la pharmacogénétique des AVK pourraient être

alors reléguées au second plan en faveur de l'utilisation de ces nouvelles molécules.

En dépit de ces considérations, la pharmacogénétique est largement amenée à jouer un rôle capital dans l'amélioration des thérapies médicamenteuses et dans la manière dont va être pratiquée la médecine de demain. Malgré la remarquable expansion du champ de la pharmacogénétique ces dernières années, elle peine aujourd'hui à s'imposer dans la pratique médicale et reste principalement cantonnée à certains secteurs, comme la cancérologie, et ce pour diverses raisons. Parmi les plus importantes, citons le faible pourcentage de molécules pour lesquelles on dispose d'une information pharmacogénétique précise et fiable ; le manque de formation des médecins à l'utilisation et à l'interprétation des données de pharmacogénétique et leur manque de conviction quant à l'intérêt de cette information ; l'absence de recommandations par les autorités compétentes de l'utilisation de tests de pharmacogénétique et le non remboursement de ces derniers, le plus souvent par manque de preuves de leur utilité sur les plans clinique et économique. Pour améliorer son intégration dans la pratique clinique, la pharmacogénétique doit surmonter un certain nombre d'obstacles d'ordre social, éthique, juridique, économique, biotechnologique et scientifique. Sur ce dernier point, la génétique des populations peut apporter une aide précieuse. En effet, la pharmacogénétique souffre le plus souvent du manque de connaissance de la valeur prédictive d'un génotype sur le phénotype, qui résulte notamment de la faible reproductibilité des associations génotype/phénotype identifiées lorsque des échantillons indépendants sont considérés, et notamment lorsqu'ils sont constitués d'individus d'origine ethnique différente. Bien qu'il y ait une réelle nécessité à mieux définir les phénotypes de la réponse aux médicaments, l'importante variabilité interindividuelle et inter-populationnelle des déterminants génétiques de la réponse aux médicaments est la cause majeure de ce problème. C'est pourquoi il est important de bien caractériser la diversité pharmacogénétique des populations humaines dans différentes régions du monde et de comprendre les mécanismes évolutifs qui la sous-tendent. C'est dans ce contexte que se positionne ce travail de thèse. En sélectionnant des

variants génétiques de pharmacogènes présentant à la fois une extrême différenciation entre les populations humaines et une probabilité élevée d'être fonctionnels, une telle approche peut permettre d'identifier des variants susceptibles d'expliquer les différences de réponse aux médicaments entre populations, même si des études fonctionnelles complémentaires sont bien-sûr nécessaires pour préciser leur rôle. Étant donné que les différences génétiques entre les populations humaines sont plus de nature quantitative (fréquence des variants) que qualitative (nature des variants), ces variants sont également susceptibles d'expliquer une bonne part de la variabilité interindividuelle de réponse aux médicaments et peuvent constituer des biomarqueurs intéressants à inclure dans des tests pharmacogénétiques pour sécuriser et rendre plus efficaces les thérapies médicamenteuses.

En outre, le développement actuel des techniques de séquençage à haut-débit donne accès dorénavant aux données de séquence du génome entier chez un grand nombre d'individus de différentes populations du monde. Ces données permettent d'étudier des variants peu représentés jusqu'alors dans les bases de données de génotypage, qu'il peut être intéressant de considérer dans des études de pharmacogénétique. En effet, les variants rares et peu fréquents sont très abondants dans le génome humain (1000 Genomes Project Consortium et al., 2012), notamment dans les pharmacogènes : une étude récente ayant séquencé 202 gènes codant des cibles pharmacologiques de médicaments chez plus de 14 000 individus a révélé que plus de 95 % des variants identifiés avaient une MAF ≤ 0,05 (Nelson et al., 2012). Par ailleurs, ces variants semblent être enrichis en variants potentiellement fonctionnels, car 85 % des variants non-synonymes et 90 % des variants entraînant un codon stop ou modifiant le site d'épissage présentent une MAF inférieure à 0,05, alors que cette proportion n'est que de 65 % pour les variants synonymes (1000 Genomes Project Consortium et al., 2012). En revanche, il serait nécessaire d'étudier la répartition de ces variants dans un panel plus large de populations que celles incluses actuellement dans le Projet 1000 Génomes. En effet l'ensemble de la diversité génétique humaine est loin d'être représentée par ces seules populations (Lu and Xu, 2013).

Un autre aspect de ce travail de thèse a consisté à rechercher des signatures génomiques de la sélection naturelle, susceptibles d'expliquer les profils de différenciation géographique atypiques observés pour certains variants de gènes d'intérêt majeur en pharmacogénétique. Des différences importantes de fréquence allélique entre populations peuvent en effet être le résultat de forces non sélectives (Hofer et al. 2009) et il est nécessaire de mettre en œuvre des tests formels de sélection pour préciser la nature des forces évolutives sous-jacentes. Si l'on peut démontrer qu'un variant présentant un profil de différenciation extrême a été la cible d'une sélection, cela conforte sa fonctionnalité et sa possible implication dans le déterminisme d'un phénotype. Étant donnée la nature des gènes étudiés (pharmacogènes), il est probable qu'il soit plus particulièrement impliqué dans la variabilité de réponse à des médicaments. Le *design* de notre stratégie de détection de la sélection mise en œuvre dans l'article 2 ne permet cependant pas de déterminer si le variant pharmacogénétique d'intérêt a été la cible directe de la sélection. Une étude plus approfondie de la signature de sélection dans une région plus étendue que celle incluant le seul gène d'intérêt, comme celle mise en œuvre dans la région génomique de *VKORC1* (partie 2 de cette thèse), est nécessaire pour localiser plus finement la cible de la sélection. Remarquons toutefois que même avec une approche très approfondie, il peut être difficile d'identifier précisément le variant sélectionné. Comme l'a montré clairement l'étude de *VKORC1*, les résultats des tests de sélection dépendent de la densité de marqueurs génétiques considérés et des tests de sélection employés. L'analyse des résultats des tests de sélection réalisés sur les données de séquence du projet 1000 Génomes a permis en effet d'affiner ceux obtenus avec les données de génotypage du panel HGDP-CEPH. Nous avons ainsi démontré que la sélection naturelle pouvait jouer sur des variants situés dans d'autres gènes (*PRSS53* dans l'exemple de *VKORC1*), entraînant avec eux les variants situés dans les pharmacogènes d'intérêt. Cela ne peut être mis en évidence que si l'on étudie le gène d'intérêt dans son contexte génomique, et sur des données de séquences. Il faut cependant rester prudent dans l'interprétation des résultats des tests de sélection actuellement disponibles, qui ne sont pas forcément très puissants pour la localisation fine de la cible

de sélection. En effet, l'identification des variants génétiques précis impliqués dans l'évolution adaptative des populations est particulièrement complexe au regard des échelles de temps considérées. Notre étude a toutefois permis de mettre en évidence le rôle majeur joué par la sélection naturelle dans la structure génétique des populations humaines pour les gènes de la réponse aux médicaments et de révéler des signatures de sélection encore non décrites à ce jour. En effet, notre approche consistant à combiner plusieurs tests de sélection complémentaires et à déterminer la signification statistique des scores calculés par l'utilisation d'une approche empirique basée sur des distributions génome entier nous permet d'être confiants dans les signatures génomiques de sélection identifiées.

A l'exception du gène *NAT2*, nous nous sommes essentiellement intéressés à la détection de la sélection positive, et plus précisément, aux événements de type balayage sélectif complet ou presque complet répondant au modèle classique « hard sweep ». Or, il est probable que d'autres types de sélection naturelle aient été à l'œuvre dans le passé sur les gènes impliqués aujourd'hui dans la réponse aux médicaments. En effet comme nous l'avons introduit dans la première partie de cette thèse, la sélection positive n'agit pas seulement *via* le mécanisme de balayage sélectif classiquement étudié, dans lequel une unique mutation avantageuse est, sous l'effet d'une forte pression de sélection positive, rapidement amenée à une fréquence élevée dans une ou plusieurs populations. Au contraire, il semblerait que ce processus sélectif soit assez rare dans l'histoire évolutive de notre espèce (Flintoft, 2011; Hernandez et al., 2011). La sélection positive agirait principalement en générant de faibles écarts de fréquence allélique, et il semble qu'elle soit plus efficace en cas de changement environnemental brusque si elle agit à partir de la variation génétique préexistante plutôt que sur de nouveaux allèles (adaptation polygénique), ou en ciblant plusieurs allèles à la fois plutôt qu'un seul (*soft sweep*) (Hermisson and Pennings, 2005; Pritchard and Di Rienzo, 2010; Pritchard et al., 2010; Ralph and Coop, 2010). Il est notamment avéré que ces mécanismes sélectifs sont intervenus au cours des phénomènes d'adaptation locale des populations à des variations de l'environnement climatique et nutritionnel (Hancock et al., 2010). Cependant

ces phénomènes laissent des marques moins nettes sur le génome humain et sont donc plus difficiles à mettre en évidence, en particulier si l'on ne dispose que de données de génotypage. Avec la plus grande disponibilité des données de séquences et le développement de nouvelles méthodes adaptées, il est cependant permis d'espérer qu'à l'avenir il sera possible de les étudier.

Il aurait été intéressant également de regarder les variants pharmacogénétiques présentant un niveau de différenciation génétique particulièrement faible, révélateur possible de la sélection balancée ou d'une autre forme de sélection favorisant le même allèle dans les différentes populations étudiées. Nous avons observé que parmi les 90 variants clés de la réponse aux médicaments, seuls deux affichent une valeur de F_{ST} inter-populationnel inhabituellement faible : rs2046934 dans *P2RY12* (*P*-values empiriques de la distribution principale et de la sous-distribution = 0,03, MAF globale = 0,153) et rs1801028 dans *DRD2* (*P*-values empiriques de la distribution principale et de la sous-distribution = 0,03, MAF globale = 0,023). Ce petit nombre de variants pourrait s'expliquer de deux manières : soit ces processus de sélection homogénéisants sont plus rares que les processus de sélection qui conduisent au contraire à une différenciation des populations, soit les méthodes que nous avons utilisées, basées sur le F_{ST}, sont moins puissantes pour les mettre en évidence. Si nous ne pouvons pas entièrement répondre à la question, nous pouvons, à partir des distributions empiriques des F_{ST} que nous avons calculées sur l'ensemble du génome, donner quelques éléments de réponse en faveur de la deuxième hypothèse. En effet, nous avons pu constater une masse en zéro des F_{ST} se traduisant par des valeurs seuils de F_{ST} très faibles, parfois même à zéro, pour les 1er et 5^{e} percentiles, rendant difficile l'identification des marqueurs avec des F_{ST} plus faibles que ces valeurs seuils.

Pour finir, il faut s'attendre dans les années à venir, à ce que le développement des technologies de nouvelle génération d'approche globale « omique » permette d'étendre le champ d'investigation de la

variabilité de réponse aux médicaments et de la sélection naturelle. Ainsi, il sera possible d'explorer les aspects moléculaire, cellulaire et tissulaire, grâce à l'analyse du transcriptome, du protéome, du génome somatique, du métabolome et de l'épigénome. Le développement de nouvelles méthodes de détection des signatures de sélection, plus adaptées aux données de séquençage et à la complexité des mécanismes sous-tendant l'histoire évolutive des gènes, en corrélation avec des données environnementales, permettra d'identifier de façon plus complète les gènes et les variants ayant joué un rôle important dans l'adaptation des populations humaines à leur environnement.

Au cours de cette thèse, nous n'avons cessé de démontrer le rôle fondamental joué par la sélection naturelle dans la différenciation pharmacogénétique des populations humaines. Cette constatation nous confirme le grand intérêt qu'il y a à détecter les pressions de sélection auxquelles ont été soumis les gènes impliqués dans la réponse aux xénobiotiques par le passé, pour faciliter l'identification des variants pharmacogénétiques responsables d'une partie de la variabilité inter-populationnelle de réponse aux médicaments aujourd'hui. L'apport essentiel de la génétique des populations à l'amélioration de la compréhension des facteurs génétiques impliqués dans les traits complexes comme la réponse aux médicaments, en fait un outil majeur au service du développement d'une médecine personnalisée, sûre et efficace pour tous les individus. Cela permet également d'apporter des clés de compréhension précieuses de l'histoire adaptative de l'homme en réponse à des variations de son environnement chimique.

Références bibliographiques

1000 Genomes Project Consortium, Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061–1073.

1000 Genomes Project Consortium, Abecasis, G.R., Auton, A., Brooks, L.D., DePristo, M.A., Durbin, R.M., Handsaker, R.E., Kang, H.M., Marth, G.T., and McVean, G.A. (2012). An integrated map of genetic variation from 1,092 human genomes. Nature *491*, 56–65.

AFSSAPS (2008). Hospitalisations dues aux effets indésirables des médicaments : résultats d'une étude nationale Point sur la nouvelle campagne d'information sur les traitements anticoagulants antivitamine K.

Agúndez, J.A.G., Martínez, C., Pérez-Sala, D., Carballo, M., Torres, M.J., and García-Martín, E. (2009). Pharmacogenomics in aspirin intolerance. Curr. Drug Metab. *10*, 998–1008.

Akey, J.M., Zhang, G., Zhang, K., Jin, L., and Shriver, M.D. (2002). Interrogating a high-density SNP map for signatures of natural selection. Genome Res *12*, 1805–1814.

Aklillu, E., Leong, C., Loebstein, R., Halkin, H., and Gak, E. (2008). VKORC1 Asp36Tyr warfarin resistance marker is common in Ethiopian individuals. Blood *111*, 3903–3904.

Almquist, H.J., and Klose, A.A. (1939). THE ANTI-HEMORRHAGIC ACTIVITY OF PURE SYNTHETIC PHTHIOCOL. J. Am. Chem. Soc. *61*, 1611–1611.

ALVING, A.S., CARSON, P.E., FLANAGAN, C.L., and ICKES, C.E. (1956). Enzymatic deficiency in primaquine-sensitive erythrocytes. Science *124*, 484–485.

Aminkeng, F., Ross, C.J.D., Rassekh, S.R., Brunham, L.R., Sistonen, J., Dube, M.-P., Ibrahim, M., Nyambo, T.B., Omar, S.A., Froment, A., et al. (2014). Higher frequency of genetic variants conferring increased risk for ADRs for commonly used drugs treating cancer, AIDS and tuberculosis in persons of African descent. Pharmacogenomics J. *14*, 160–170.

An, H.-R., Wu, X.-Q., Wang, Z.-Y., Zhang, J.-X., and Liang, Y. (2012). NAT2 and CYP2E1 polymorphisms associated with antituberculosis drug-induced hepatotoxicity in Chinese patients. Clin. Exp. Pharmacol. Physiol. *39*, 535–543.

ANSM (2012). Les anticoagulants en France en 2012 : Etat des lieux et surveillance.

Apretna, E., Haramburu, F., Taboulet, F., and Bégaud, B. (2005). [Medical and socio-economical impact of drug-induced adverse reactions]. Presse Médicale Paris Fr. 1983 *34*, 271–276.

Auton, A., Bryc, K., Boyko, A.R., Lohmueller, K.E., Novembre, J., Reynolds, A., Indap, A., Wright, M.H., Degenhardt, J.D., Gutenkunst, R.N., et al. (2009). Global distribution of genomic diversity underscores rich complex history of continental human populations. Genome Res. *19*, 795–803.

Bains, R.K., Kovacevic, M., Plaster, C.A., Tarekegn, A., Bekele, E., Bradman, N.N., and Thomas, M.G. (2013). Molecular diversity and population structure at the Cytochrome P450 3A5 gene in Africa. BMC Genet. *14*, 34.

Balram, C., Sabapathy, K., Fei, G., Khoo, K.S., and Lee, E.J.D. (2002). Genetic polymorphisms of UDP-glucuronosyltransferase in Asians: UGT1A1*28 is a common allele in Indians. Pharmacogenetics *12*, 81–83.

Bamshad, M., and Wooding, S.P. (2003). Signatures of natural selection in the human genome. Nat. Rev. Genet. *4*, 99–111.

Barbujani, G., and Colonna, V. (2010). Human genome diversity: frequently asked questions. Trends Genet. TIG *26*, 285–295.

Barreiro, L.B., and Quintana-Murci, L. (2010). From evolutionary genetics to human immunology: how selection shapes host defence genes. Nat. Rev. Genet. *11*, 17–30.

Barreiro, L.B., Laval, G., Quach, H., Patin, E., and Quintana-Murci, L. (2008). Natural selection has driven population differentiation in modern humans. Nat Genet *40*, 340–345.

BARRITT, D.W., and JORDAN, S.C. (1960). Anticoagulant drugs in the treatment of pulmonary embolism. A controlled trial. Lancet *1*, 1309–1312.

Bates, D.W., Cullen, D.J., Laird, N., Petersen, L.A., Small, S.D., Servi, D., Laffel, G., Sweitzer, B.J., Shea, B.F., and Hallisey, R. (1995). Incidence of adverse drug events and potential adverse drug events. Implications for prevention. ADE Prevention Study Group. JAMA J. Am. Med. Assoc. *274*, 29–34.

Beaumont, M.A., and Nichols, R.A. (1996). Evaluating loci for use in the genetic analysis of population structure. R. Soc. *263*, 1619–1626.

Bégaud, B., and Costagliola, D. (2013). Rapport sur la surveillance et la promotion du bon usage du médicament en France.

Bell, R.G. (1978). Metabolism of vitamin K and prothrombin synthesis: anticoagulants and the vitamin K--epoxide cycle. Fed. Proc. *37*, 2599–2604.

Bennett, J.W., Pybus, B.S., Yadava, A., Tosh, D., Sousa, J.C., McCarthy, W.F., Deye, G., Melendez, V., and Ockenhouse, C.F. (2013). Primaquine failure and cytochrome P-450 2D6 in Plasmodium vivax malaria. N. Engl. J. Med. *369*, 1381–1382.

Berkner, K.L., and Runge, K.W. (2004). The physiology of vitamin K nutriture and vitamin K-dependent protein function in atherosclerosis. J Thromb Haemost *2*, 2118–2132.

Bersaglieri, T., Sabeti, P.C., Patterson, N., Vanderploeg, T., Schaffner, S.F., Drake, J.A., Rhodes, M., Reich, D.E., and Hirschhorn, J.N. (2004). Genetic signatures of strong recent positive selection at the lactase gene. Am J Hum Genet *74*, 1111–1120.

Beulens, J.W.J., Booth, S.L., van den Heuvel, E.G.H.M., Stoecklin, E., Baka, A., and Vermeer, C. (2013). The role of menaquinones (vitamin K_2) in human health. Br. J. Nutr. *110*, 1357–1368.

Beutler, E., Gelbart, T., and Demina, A. (1998). Racial variability in the UDP-glucuronosyltransferase 1 (UGT1A1) promoter: a balanced polymorphism for regulation of bilirubin metabolism? Proc. Natl. Acad. Sci. U. S. A. *95*, 8170–8174.

Bigham, A., Bauchet, M., Pinto, D., Mao, X., Akey, J.M., Mei, R., Scherer, S.W., Julian, C.G., Wilson, M.J., López Herráez, D., et al. (2010). Identifying signatures of natural selection in Tibetan and Andean populations using dense genome scan data. PLoS Genet. *6*, e1001116.

Bigham, A.W., Mao, X., Mei, R., Brutsaert, T., Wilson, M.J., Julian, C.G., Parra, E.J., Akey, J.M., Moore, L.G., and Shriver, M.D. (2009). Identifying positive selection candidate loci for high-altitude adaptation in Andean populations. Hum. Genomics *4*, 79–90.

Binkley, S.B., McKee, R.W., Thayer, S.A., and Doisy, E.A. (1940). The constitution of vitamin K2. J. Biol. Chem. 721–729.

Biomarkers Definitions Working Group. (2001). Biomarkers and surrogate endpoints: preferred definitions and conceptual framework. Clin. Pharmacol. Ther. *69*, 89–95.

Bloche, M.G. (2006). Race, money and medicines. J. Law Med. Ethics J. Am. Soc. Law Med. Ethics *34*, 555–558, 480.

Blum, M., Demierre, A., Grant, D.M., Heim, M., and Meyer, U.A. (1991). Molecular mechanism of slow acetylation of drugs and carcinogens in humans. Proc. Natl. Acad. Sci. U. S. A. *88*, 5237–5241.

Bodin, L., Verstuyft, C., Tregouet, D.-A., Robert, A., Dubert, L., Funck-Brentano, C., Jaillon, P., Beaune, P., Laurent-Puig, P., Becquemont, L., et al. (2005). Cytochrome P450 2C9 (CYP2C9) and vitamin K epoxide reductase (VKORC1) genotypes as determinants of acenocoumarol sensitivity. Blood *106*, 135–140.

BONICKE, R., and REIF, W. (1953). [Enzymatic inactivation of isonicotinic acid hydrizide in human and animal organism]. Naunyn-Schmiedebergs Arch. Für Exp. Pathol. Pharmakol. *220*, 321–323.

Booth, S.L., and Suttie, J.W. (1998). Dietary intake and adequacy of vitamin K. J. Nutr. *128*, 785–788.

Borobia, A.M., Lubomirov, R., Ramírez, E., Lorenzo, A., Campos, A., Muñoz-Romo, R., Fernández-Capitán, C., Frías, J., and Carcas, A.J. (2012). An acenocoumarol dosing algorithm using clinical and pharmacogenetic data in Spanish patients with thromboembolic disease. PloS One *7*, e41360.

Bradford, L.D., and Kirlin, W.G. (1998). Polymorphism of CYP2D6 in Black populations: implications for psychopharmacology. Int. J. Neuropsychopharmacol. Off. Sci. J. Coll. Int. Neuropsychopharmacol. CINP *1*, 173–185.

Budnitz, D.S., Shehab, N., Kegler, S.R., and Richards, C.L. (2007). Medication use leading to emergency department visits for adverse drug events in older adults. Ann. Intern. Med. *147*, 755–765.

Bügel, S. (2008). Vitamin K and bone health in adult humans. Vitam. Horm. *78*, 393–416.

Burnett, A., Tiongson, J., Downey, R., and Mahan, C.E. (2013). The hidden costs of anticoagulation in hospitalized patients with non-valvular atrial fibrillation. Expert Opin. Pharmacother. *14*, 1119–1133.

Butcher, N.J., Boukouvala, S., Sim, E., and Minchin, R.F. (2002). Pharmacogenetics of the arylamine N-acetyltransferases. Pharmacogenomics J. *2*, 30–42.

Cal, S., Peinado, J.R., Llamazares, M., Quesada, V., Moncada-Pazos, A., Garabaya, C., and López-Otín, C. (2006). Identification and characterization of human polyserase-3, a novel protein with tandem serine-protease domains in the same polypeptide chain. BMC Biochem. *7*, 9.

Cann, R.L. (2001). Genetic clues to dispersal in human populations: retracing the past from the present. Science *291*, 1742–1748.

Cann, H.M., de Toma, C., Cazes, L., Legrand, M.F., Morel, V., Piouffre, L., Bodmer, J., Bodmer, W.F., Bonne-Tamir, B., Cambon-Thomsen, A., et al. (2002). A human genome diversity cell line panel. Science *296*, 261–262.

Cann, R.L., Stoneking, M., and Wilson, A.C. (1987). Mitochondrial DNA and human evolution. Nature *325*, 31–36.

Cappellini, M.D., and Fiorelli, G. (2008). Glucose-6-phosphate dehydrogenase deficiency. Lancet *371*, 64–74.

Carlberg, C., Seuter, S., de Mello, V.D.F., Schwab, U., Voutilainen, S., Pulkki, K., Nurmi, T., Virtanen, J., Tuomainen, T.-P., and Uusitupa, M. (2013). Primary Vitamin D Target Genes Allow a Categorization of Possible Benefits of Vitamin D3 Supplementation. PLoS ONE *8*, e71042.

Carlson, C.S., Thomas, D.J., Eberle, M.A., Swanson, J.E., Livingston, R.J., Rieder, M.J., and Nickerson, D.A. (2005). Genomic regions exhibiting positive selection identified from dense genotype data. Genome Res *15*, 1553–1565.

Casto, A.M., and Feldman, M.W. (2011). Genome-wide association study SNPs in the human genome diversity project populations: does selection affect unlinked SNPs with shared trait associations? PLoS Genet. *7*, e1001266.

Cavalli-Sforza, L.L., and Feldman, M.W. (2003). The application of molecular genetic approaches to the study of human evolution. Nat. Genet. *33 Suppl*, 266–275.

Celorrio, D., Bujanda, L., Chbel, F., Sánchez, D., Martinez-Jarreta, B., and de Pancorbo, M.M. (2011). Alcohol-metabolizing enzyme gene polymorphisms in the Basque Country, Morocco, and Ecuador. Alcohol. Clin. Exp. Res. *35*, 879–884.

Cha, P.C., Mushiroda, T., Takahashi, A., Kubo, M., Minami, S., Kamatani, N., and Nakamura, Y. (2010). Genome-wide association study identifies genetic determinants of warfarin responsiveness for Japanese. Hum Mol Genet *19*, 4735–4744.

Chaix, R., Quintana-Murci, L., Hegay, T., Hammer, M.F., Mobasher, Z., Austerlitz, F., and Heyer, E. (2007). From social to genetic structures in central Asia. Curr. Biol. CB *17*, 43–48.

Chaix, R., Cao, C., and Donnelly, P. (2008). Is mate choice in humans MHC-dependent? PLoS Genet. *4*, e1000184.

Charlesworth, B., Morgan, M.T., and Charlesworth, D. (1993). The effect of deleterious mutations on neutral molecular variation. Genetics *134*, 1289–1303.

Chen, H., Patterson, N., and Reich, D. (2010a). Population differentiation as a test for selective sweeps. Genome Res *20*, 393–402.

Chen, J., Teo, Y.Y., Toh, D.S.L., and Sung, C. (2010b). Interethnic comparisons of important pharmacology genes using SNP databases: potential application to drug regulatory assessments. Pharmacogenomics *11*, 1077–1094.

Chowbay, B., Zhou, D.S., and Lee, E.J.D. (2008). An Interethnic Comparison of Polymorphisms of the Genes Encoding Drug-Metabolizing Enzymes and Drug Transporters: Experience in Singapore.

Chung, W.-H., Hung, S.-I., Hong, H.-S., Hsih, M.-S., Yang, L.-C., Ho, H.-C., Wu, J.-Y., and Chen, Y.-T. (2004). Medical genetics: a marker for Stevens-Johnson syndrome. Nature *428*, 486.

Chung, W.-H., Hung, S.-I., and Chen, Y.-T. (2010). Genetic predisposition of life-threatening antiepileptic-induced skin reactions. Expert Opin. Drug Saf. *9*, 15–21.

Clark, A.G., Hubisz, M.J., Bustamante, C.D., Williamson, S.H., and Nielsen, R. (2005). Ascertainment bias in studies of human genome-wide polymorphism. Genome Res *15*, 1496–1502.

Classen, D.C., Pestotnik, S.L., Evans, R.S., Lloyd, J.F., and Burke, J.P. (1997). Adverse drug events in hospitalized patients. Excess length of stay, extra costs, and attributable mortality. JAMA J. Am. Med. Assoc. *277*, 301–306.

CLAYMAN, C.B., ARNOLD, J., HOCKWALD, R.S., YOUNT, E.H., Jr, EDGCOMB, J.H., and ALVING, A.S. (1952). Toxicity of primaquine in Caucasians. J. Am. Med. Assoc. *149*, 1563–1568.

Cohn, J.N., Johnson, G., Ziesche, S., Cobb, F., Francis, G., Tristani, F., Smith, R., Dunkman, W.B., Loeb, H., and Wong, M. (1991). A comparison of enalapril with hydralazine-isosorbide dinitrate in the treatment of chronic congestive heart failure. N. Engl. J. Med. *325*, 303–310.

Conrad, D.F., Jakobsson, M., Coop, G., Wen, X., Wall, J.D., Rosenberg, N.A., and Pritchard, J.K. (2006). A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. Nat Genet *38*, 1251–1260.

Cooper, G.M., Johnson, J.A., Langaee, T.Y., Feng, H., Stanaway, I.B., Schwarz, U.I., Ritchie, M.D., Stein, C.M., Roden, D.M., Smith, J.D., et al. (2008). A genome-wide scan for common genetic variants with a large influence on warfarin maintenance dose. Blood *112*, 1022–1027.

Cordero, P., and Ashley, E.A. (2012). Whole-genome sequencing in personalized therapeutics. Clin. Pharmacol. Ther. *91*, 1001–1009.

D'Andrea, G., D'Ambrosio, R.L., Di Perna, P., Chetta, M., Santacroce, R., Brancaccio, V., Grandone, E., and Margaglione, M. (2005). A polymorphism in the VKORC1 gene is associated with an interindividual variability in the dose-anticoagulant effect of warfarin. Blood *105*, 645–649.

D'Andrea, G., D'Ambrosio, R., and Margaglione, M. (2008). Oral anticoagulants: Pharmacogenetics Relationship between genetic and non-genetic factors. Blood Rev *22*, 127–140.

Daly, A.K. (2010). Genome-wide association studies in pharmacogenomics. Nat. Rev. Genet. *11*, 241–246.

Daly, A.K., and King, B.P. (2003). Pharmacogenetics of oral anticoagulants. Pharmacogenetics *13*, 247–252.

Dam, H. (1935). The antihaemorrhagic vitamin of the chick. Biochem. J. *29*, 1273–1285.

Dam, H., and Schönheyder, F. (1934). A deficiency disease in chicks resembling scurvy. Biochem. J. *28*, 1355–1359.

Dandara, C., Swart, M., Mpeta, B., Wonkam, A., and Masimirembwa, C. (2014). Cytochrome P450 pharmacogenetics in African populations: implications for public health. Expert Opin. Drug Metab. Toxicol.

Darwin, C. (1859). On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. Nature.

Debray, M., Pautas, E., Couturier, P., Franco, A., and Siguret, V. (2003). [Oral anticoagulants in the elderly]. Rev. Médecine Interne Fondée Par Société Natl. Francaise Médecine Interne *24*, 107–117.

Delser, P.M., and Fuselli, S. (2013). Human loci involved in drug biotransformation: worldwide genetic variation, population structure, and pharmacogenetic implications. Hum. Genet. *132*, 563–577.

DeLuca, H.F. (2004). Overview of general physiologic features and functions of vitamin D. Am. J. Clin. Nutr. *80*, 1689S–1696S.

Denison, M.S., Pandini, A., Nagy, S.R., Baldwin, E.P., and Bonati, L. (2002). Ligand binding and activation of the Ah receptor. Chem. Biol. Interact. *141*, 3–24.

Dimasi, J.A. (2001). Risks in new drug development: approval success rates for investigational drugs. Clin. Pharmacol. Ther. *69*, 297–307.

DOUGLAS, A.S. (1955). Mode of action of coumarin drugs. Br. Med. Bull. *11*, 39–44.

Duran-Frigola, M., and Aloy, P. (2013). Analysis of chemical and biological features yields mechanistic insights into drug side effects. Chem. Biol. *20*, 594–603.

Duster, T. (2007). Medicalisation of race. The Lancet *369*, 702–704.

Ebbesen, J., Buajordet, I., Erikssen, J., Brørs, O., Hilberg, T., Svaar, H., and Sandvik, L. (2001). Drug-related deaths in a department of internal medicine. Arch. Intern. Med. *161*, 2317–2323.

Edenberg, H.J. (2000). Regulation of the mammalian alcohol dehydrogenase genes. Prog. Nucleic Acid Res. Mol. Biol. *64*, 295–341.

Elder, S.J., Haytowitz, D.B., Howe, J., Peterson, J.W., and Booth, S.L. (2006). Vitamin k contents of meat, dairy, and fast food in the u.s. Diet. J. Agric. Food Chem. *54*, 463–467.

Enard, W., Przeworski, M., Fisher, S.E., Lai, C.S.L., Wiebe, V., Kitano, T., Monaco, A.P., and Pääbo, S. (2002). Molecular evolution of FOXP2, a gene involved in speech and language. Nature *418*, 869–872.

Evans, D.A. (1989). N-acetyltransferase. Pharmacol. Ther. *42*, 157–234.

EVANS, D.A., MANLEY, K.A., and McKUSICK, V.A. (1960). Genetic control of isoniazid metabolism in man. Br. Med. J. *2*, 485–491.

Fay, J.C., and Wu, C.I. (2000). Hitchhiking under positive Darwinian selection. Genetics *155*, 1405–1413.

FDA (2005). Hepatic dysfunction, pancreatitis, UGT1A1. Final Label FDA approval drug label.

Feng, Z., Smith, D.L., McKenzie, F.E., and Levin, S.A. (2004). Coupling ecology and evolution: malaria and the S-gene across time scales. Math. Biosci. *189*, 1–19.

Ferder, N.S., Eby, C.S., Deych, E., Harris, J.K., Ridker, P.M., Milligan, P.E., Goldhaber, S.Z., King, C.R., Giri, T., McLeod, H.L., et al. (2010). Ability of VKORC1 and CYP2C9 to predict therapeutic warfarin dose during the initial weeks of therapy. J. Thromb. Haemost. JTH *8*, 95–100.

Ferguson, R.J., Doll, M.A., Rustan, T.D., Gray, K., and Hein, D.W. (1994). Cloning, expression, and functional characterization of two mutant (NAT2(191) and NAT2(341/803)) and wild-type human polymorphic N-acetyltransferase (NAT2) alleles. Drug Metab. Dispos. Biol. Fate Chem. *22*, 371–376.

Ferland, G. (2012). The discovery of vitamin K and its clinical applications. Ann. Nutr. Metab. *61*, 213–218.

Fernandes, M.R., de Carvalho, D.C., dos Santos, Â.K.C.R., dos Santos, S.E.B., de Assumpção, P.P., Burbano, R.M.R., and dos Santos, N.P.C. (2013). Association of slow acetylation profile of NAT2 with breast and gastric cancer risk in Brazil. Anticancer Res. *33*, 3683–3689.

Ferrell, P.B., Jr, and McLeod, H.L. (2008). Carbamazepine, HLA-B*1502 and risk of Stevens-Johnson syndrome and toxic epidermal necrolysis: US FDA recommendations. Pharmacogenomics *9*, 1543–1546.

Fieser, L.F. (1939). Synthesis of Vitamin K1. J. Am. Chem. Soc. *61*, 3467–3475.

Fisher, S.E., Vargha-Khadem, F., Watkins, K.E., Monaco, A.P., and Pembrey, M.E. (1998). Localisation of a gene implicated in a severe speech and language disorder. Nat. Genet. *18*, 168–170.

Flintoft, L. (2011). Human evolution: Sweep model is swept away. Nat. Rev. Genet. *12*, 228–229.

Fox, A.L. (1932). The Relationship between Chemical Constitution and Taste. Proc. Natl. Acad. Sci. U. S. A. *18*, 115–120.

Franciotta, D., Kwan, P., and Perucca, E. (2009). Genetic basis for idiosyncratic reactions to antiepileptic drugs. Curr. Opin. Neurol. *22*, 144–149.

Fretland, A.J., Leff, M.A., Doll, M.A., and Hein, D.W. (2001). Functional characterization of human N-acetyltransferase 2 (NAT2) single nucleotide polymorphisms. Pharmacogenetics *11*, 207–215.

Fu, Y.X. (1997). Statistical tests of neutrality of mutations against population growth, hitchhiking and background selection. Genetics *147*, 915–925.

Fu, Y.X., and Li, W.H. (1993). Statistical tests of neutrality of mutations. Genetics *133*, 693–709.

Fumagalli, M., Pozzoli, U., Cagliani, R., Comi, G.P., Riva, S., Clerici, M., Bresolin, N., and Sironi, M. (2009). Parasites represent a major selective force for interleukin genes and shape the genetic predisposition to autoimmune conditions. J. Exp. Med. *206*, 1395–1408.

Fuselli, S., de Filippo, C., Mona, S., Sistonen, J., Fariselli, P., Destro-Bisol, G., Barbujani, G., Bertorelle, G., and Sajantila, A. (2010). Evolution of detoxifying systems: the role of environment and population history in shaping genetic diversity at human CYP2D6 locus. Pharmacogenet. Genomics *20*, 485–499.

Gabriel, S.B., Schaffner, S.F., Nguyen, H., Moore, J.M., Roy, J., Blumenstiel, B., Higgins, J., DeFelice, M., Lochner, A., Faggart, M., et al. (2002). The structure of haplotype blocks in the human genome. Science *296*, 2225–2229.

Gage, B.F., Eby, C., Milligan, P.E., Banet, G.A., Duncan, J.R., and McLeod, H.L. (2004). Use of pharmacogenetics and clinical factors to predict the maintenance dose of warfarin. Thromb. Haemost. *91*, 87–94.

Gage, B.F., Eby, C., Johnson, J.A., Deych, E., Rieder, M.J., Ridker, P.M., Milligan, P.E., Grice, G., Lenzini, P., Rettie, A.E., et al. (2008). Use of pharmacogenetic and clinical factors to predict the therapeutic dose of warfarin. Clin Pharmacol Ther *84*, 326–331.

Gamazon, E.R., Duan, S., Zhang, W., Huang, R.S., Kistner, E.O., Dolan, M.E., and Cox, N.J. (2010). PACdb: a database for cell-based pharmacogenomics. Pharmacogenet. Genomics *20*, 269–273.

Gamazon, E.R., Huang, R.S., Dolan, M.E., and Cox, N.J. (2011). Copy number polymorphisms and anticancer pharmacogenomics. Genome Biol. *12*, R46.

Gamazon, E.R., Skol, A.D., and Perera, M.A. (2012). The limits of genome-wide methods for pharmacogenomic testing. Pharmacogenet. Genomics *22*, 261–272.

Gardner, M., Bertranpetit, J., and Comas, D. (2008). Worldwide genetic variation in dopamine and serotonin pathway genes: implications for association studies. Am. J. Med. Genet. Part B Neuropsychiatr. Genet. Off. Publ. Int. Soc. Psychiatr. Genet. *147B*, 1070–1075.

Garrod, A.E. (2002). The incidence of alkaptonuria: a study in chemical individuality. 1902 [classical article]. Yale J. Biol. Med. *75*, 221–231.

Gemma, S., Vichi, S., and Testai, E. (2006). Individual susceptibility and alcohol effects:biochemical and genetic aspects. Ann. DellIstituto Super. Sanità *42*, 8–16.

Goldman, D., and Enoch, M.A. (1990). Genetic epidemiology of ethanol metabolic enzymes: a role for selection. World Rev. Nutr. Diet. *63*, 143–160.

Goldstein, D.B., Tate, S.K., and Sisodiya, S.M. (2003). Pharmacogenetics goes genomic. Nat. Rev. Genet. *4*, 937–947.

Grant, D.M., Tang, B.K., and Kalow, W. (1984). A simple test for acetylator phenotype using caffeine. Br. J. Clin. Pharmacol. *17*, 459–464.

Gravel, S., Henn, B.M., Gutenkunst, R.N., Indap, A.R., Marth, G.T., Clark, A.G., Yu, F., Gibbs, R.A., 1000 Genomes Project, and Bustamante, C.D. (2011). Demographic history and rare allele sharing among human populations. Proc. Natl. Acad. Sci. U. S. A. *108*, 11983–11988.

Grossman, S.R., Shlyakhter, I., Shylakhter, I., Karlsson, E.K., Byrne, E.H., Morales, S., Frieden, G., Hostetter, E., Angelino, E., Garber, M., et al. (2010). A composite of multiple signals distinguishes causal variants in regions of positive selection. Science *327*, 883–886.

Grossman, S.R., Andersen, K.G., Shlyakhter, I., Tabrizi, S., Winnicki, S., Yen, A., Park, D.J., Griesemer, D., Karlsson, E.K., Wong, S.H., et al. (2013). Identifying recent adaptations in large-scale genomic data. Cell *152*, 703–713.

Gyllensten, H., Hakkarainen, K.M., Hägg, S., Carlsten, A., Petzold, M., Rehnberg, C., and Jönsson, A.K. (2014). Economic impact of adverse drug events - a retrospective population-based cohort study of 4970 adults. PloS One *9*, e92061.

Hachad, H., Overby, C.L., Argon, S., Yeung, C.K., Ragueneau-Majlessi, I., and Levy, R.H. (2011). e-PKGene: a knowledge-based research tool for analysing the impact of genetics on drug exposure. Hum. Genomics *5*, 506–515.

Haile, D.B., Ayen, W.Y., and Tiwari, P. (2013). Prevalence and assessment of factors contributing to adverse drug reactions in wards of a tertiary care hospital, India. Ethiop. J. Health Sci. *23*, 39–48.

Hall, D., Ybazeta, G., Destro-Bisol, G., Petzl-Erler, M.L., and Di Rienzo, A. (1999). Variability at the uridine diphosphate glucuronosyltransferase 1A1 promoter in human populations and primates. Pharmacogenetics *9*, 591–599.

Hamburg, M.A., and Collins, F.S. (2010). The path to personalized medicine. N. Engl. J. Med. *363*, 301–304.

Han, Y., Gu, S., Oota, H., Osier, M.V., Pakstis, A.J., Speed, W.C., Kidd, J.R., and Kidd, K.K. (2007). Evidence of positive selection on a class I ADH locus. Am J Hum Genet *80*, 441–456.

Hancock, A.M., Witonsky, D.B., Ehler, E., Alkorta-Aranburu, G., Beall, C., Gebremedhin, A., Sukernik, R., Utermann, G., Pritchard, J., Coop, G., et al. (2010). Colloquium paper: human adaptations to diet, subsistence, and ecoregion are due to subtle shifts in allele frequency. Proc. Natl. Acad. Sci. U. S. A. *107 Suppl 2*, 8924–8930.

Harris, E.E., and Hey, J. (1999). X chromosome evidence for ancient human histories. Proc. Natl. Acad. Sci. U. S. A. *96*, 3320–3324.

HAS (2007). Fibrillation auriculaire.

Hasin, D.S., and Grant, B.F. (2004). The co-occurrence of DSM-IV alcohol abuse in DSM-IV alcohol dependence: results of the National Epidemiologic Survey on Alcohol and Related Conditions on heterogeneity that differ by population subgroup. Arch. Gen. Psychiatry *61*, 891–896.

Hebbring, S.J., Moyer, A.M., and Weinshilboum, R.M. (2008). Sulfotransferase gene copy number variation: pharmacogenetics and function. Cytogenet. Genome Res. *123*, 205–210.

Hein, D.W. (2006). N-acetyltransferase 2 genetic polymorphism: effects of carcinogen and haplotype on urinary bladder cancer risk. Oncogene *25*, 1649–1658.

Hein, D.W., Ferguson, R.J., Doll, M.A., Rustan, T.D., and Gray, K. (1994). Molecular genetics of human polymorphic N-acetyltransferase: enzymatic analysis of 15 recombinant wild-type, mutant, and chimeric NAT2 allozymes. Hum. Mol. Genet. *3*, 729–734.

Henn, B.M., Cavalli-Sforza, L.L., and Feldman, M.W. (2012). The great human expansion. Proc. Natl. Acad. Sci. U. S. A. *109*, 17758–17764.

Hermisson, J., and Pennings, P.S. (2005). Soft Sweeps Molecular Population Genetics of Adaptation From Standing Genetic Variation. Genetics *169*, 2335–2352.

Hernandez, R.D., Kelley, J.L., Elyashiv, E., Melton, S.C., Auton, A., McVean, G., 1000 Genomes Project, Sella, G., and Przeworski, M. (2011). Classic selective sweeps were rare in recent human evolution. Science *331*, 920–924.

Hernandez-Boussard, T., Whirl-Carrillo, M., Hebert, J.M., Gong, L., Owen, R., Gong, M., Gor, W., Liu, F., Truong, C., Whaley, R., et al. (2008). The pharmacogenetics and pharmacogenomics knowledge base: accentuating the knowledge. Nucleic Acids Res. *36*, D913–D918.

Hesse, L.M., He, P., Krishnaswamy, S., Hao, Q., Hogan, K., von Moltke, L.L., Greenblatt, D.J., and Court, M.H. (2004). Pharmacogenetic determinants of interindividual variability in bupropion hydroxylation by cytochrome P450 2B6 in human liver microsomes. Pharmacogenetics *14*, 225–238.

Hetherington, S.L., Singh, R.K., Lodwick, D., Thompson, J.R., Goodall, A.H., and Samani, N.J. (2005). Dimorphism in the P2Y1 ADP receptor gene is associated with increased platelet activation response to ADP. Arterioscler. Thromb. Vasc. Biol. *25*, 252–257.

Hewett, M., Oliver, D.E., Rubin, D.L., Easton, K.L., Stuart, J.M., Altman, R.B., and Klein, T.E. (2002). PharmGKB: the Pharmacogenetics Knowledge Base. Nucleic Acids Res. *30*, 163–165.

Hill, W.G. (1968). Linkage disequilibrium in finite populations. Theor. Appl. Genet. 226–231.

Hinds, D.A., Stuve, L.L., Nilsen, G.B., Halperin, E., Eskin, E., Ballinger, D.G., Frazer, K.A., and Cox, D.R. (2005). Whole-genome patterns of common DNA variation in three human populations. Science *307*, 1072–1079.

Hirsh, J., Fuster, V., Ansell, J., Halperin, J.L., and American Heart Association/American College of Cardiology Foundation (2003). American Heart Association/American College of Cardiology Foundation guide to warfarin therapy. J. Am. Coll. Cardiol. *41*, 1633–1652.

Holbrook, A.M., Pereira, J.A., Labiris, R., McDonald, H., Douketis, J.D., Crowther, M., and Wells, P.S. (2005). Systematic overview of warfarin and its drug and food interactions. Arch. Intern. Med. *165*, 1095–1106.

Holm, S. (2008). Pharmacogenetics, race and global injustice. Dev. World Bioeth. *8*, 82–88.

Hoover, E.L. (2007). There is no scientific rationale for race-based research. J. Natl. Med. Assoc. *99*, 690–692.

Huang, S.-M., and Temple, R. (2008). Is this the drug or dose for you? Impact and consideration of ethnic factors in global drug development, regulatory review, and clinical practice. Clin. Pharmacol. Ther. *84*, 287–294.

Huang, Y.-S., Chern, H.-D., Su, W.-J., Wu, J.-C., Lai, S.-L., Yang, S.-Y., Chang, F.-Y., and Lee, S.-D. (2002). Polymorphism of the N-acetyltransferase 2 gene as a susceptibility risk factor for antituberculosis drug-induced hepatitis. Hepatol. Baltim. Md *35*, 883–889.

Huebner, C., and LINK, K. (1941). Studies on the hemorrhagic sweet clover disease: VI. The synthesis of the delta-diketone derived from the hemorrhagic agent through alkaline degradation. J Biol Chem 529–534.

HUGHES, H.B., BIEHL, J.P., JONES, A.P., and SCHMIDT, L.H. (1954). Metabolism of isoniazid in man as related to the occurrence of peripheral neuritis. Am. Rev. Tuberc. *70*, 266–273.

Hylek, E.M., Go, A.S., Chang, Y., Jensvold, N.G., Henault, L.E., Selby, J.V., and Singer, D.E. (2003). Effect of intensity of oral anticoagulation on stroke severity and mortality in atrial fibrillation. N. Engl. J. Med. *349*, 1019–1026.

Indian Genome Variation Consortium (2008). Genetic landscape of the people of India: a canvas for disease gene exploration. J. Genet. *87*, 3–20.

Ingelman-Sundberg, M. (2005). Genetic polymorphisms of cytochrome P450 2D6 (CYP2D6): clinical consequences, evolutionary aspects and functional diversity. Pharmacogenomics J. *5*, 6–13.

Ingman, M., Kaessmann, H., Pääbo, S., and Gyllensten, U. (2000). Mitochondrial genome variation and the origin of modern humans. Nature *408*, 708–713.

Innocenti, F., Kroetz, D.L., Schuetz, E., Dolan, M.E., Ramírez, J., Relling, M., Chen, P., Das, S., Rosner, G.L., and Ratain, M.J. (2009). Comprehensive pharmacogenetic analysis of irinotecan neutropenia and pharmacokinetics. J. Clin. Oncol. Off. J. Am. Soc. Clin. Oncol. *27*, 2604–2614.

International HapMap 3 Consortium, Altshuler, D.M., Gibbs, R.A., Peltonen, L., Altshuler, D.M., Gibbs, R.A., Peltonen, L., Dermitzakis, E., Schaffner, S.F., Yu, F., et al. (2010). Integrating common and rare genetic variation in diverse human populations. Nature *467*, 52–58.

International HapMap Consortium (2003). The International HapMap Project. Nature *426*, 789–796.

International HapMap Consortium (2005). A haplotype map of the human genome. Nature *437*, 1299–1320.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., et al. (2007a). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

International HapMap Consortium, Frazer, K.A., Ballinger, D.G., Cox, D.R., Hinds, D.A., Stuve, L.L., Gibbs, R.A., Belmont, J.W., Boudreau, A., Hardenbol, P., et al. (2007b). A second generation human haplotype map of over 3.1 million SNPs. Nature *449*, 851–861.

Iwai, M., Suzuki, H., Ieiri, I., Otsubo, K., and Sugiyama, Y. (2004). Functional analysis of single nucleotide polymorphisms of hepatic organic anion transporter OATP1B1 (OATP-C). Pharmacogenetics *14*, 749–757.

Iyer, L., Das, S., Janisch, L., Wen, M., Ramírez, J., Karrison, T., Fleming, G.F., Vokes, E.E., Schilsky, R.L., and Ratain, M.J. (2002). UGT1A1*28 polymorphism as a determinant of irinotecan disposition and toxicity. Pharmacogenomics J. *2*, 43–47.

Jablonski, N.G., and Chaplin, G. (2010). Colloquium paper: human skin pigmentation as an adaptation to UV radiation. Proc. Natl. Acad. Sci. U. S. A. *107 Suppl 2*, 8962–8968.

Jakobsson, M., Scholz, S.W., Scheet, P., Gibbs, J.R., VanLiere, J.M., Fung, H.-C., Szpiech, Z.A., Degnan, J.H., Wang, K., Guerreiro, R., et al. (2008). Genotype, haplotype and copy-number variation in worldwide human populations. Nature *451*, 998–1003.

Johansson, I., and Ingelman-Sundberg, M. (2008). CNVs of human genes and their implication in pharmacogenetics. Cytogenet. Genome Res. *123*, 195–204.

Johnson, J.A., Gong, L., Whirl-Carrillo, M., Gage, B.F., Scott, S.A., Stein, C.M., Anderson, J.L., Kimmel, S.E., Lee, M.T.M., Pirmohamed, M., et al. (2011). Clinical Pharmacogenetics Implementation Consortium Guidelines for CYP2C9 and VKORC1 genotypes and warfarin dosing. Clin. Pharmacol. Ther. *90*, 625–629.

Jorde, L.B., Rogers, A.R., Bamshad, M., Watkins, W.S., Krakowiak, P., Sung, S., Kere, J., and Harpending, H.C. (1997). Microsatellite diversity and the demographic history of modern humans. Proc. Natl. Acad. Sci. U. S. A. *94*, 3100–3103.

Jorge, L.F., Eichelbaum, M., Griese, E.U., Inaba, T., and Arias, T.D. (1999). Comparative evolutionary pharmacogenetics of CYP2D6 in Ngawbe and Embera Amerindians of Panama and Colombia: role of selection versus drift in world populations. Pharmacogenetics 9, 217–228.

Kaessmann, H., Wiebe, V., Weiss, G., and Pääbo, S. (2001). Great ape DNA sequences reveal a reduced diversity and an expansion in humans. Nat. Genet. *27*, 155–156.

Kalow, W. (1962). Pharmacogenetics: heredity and the response to drugs. W B Saunders Co Phila. USA.

Kalow, W. (1982). Ethnic differences in drug metabolism. Clin. Pharmacokinet. *7*, 373–400.

KALOW, W., and STARON, N. (1957). On distribution and inheritance of atypical forms of human serum cholinesterase, as indicated by dibucaine numbers. Can. J. Biochem. Physiol. *35*, 1305–1320.

Kamali, F., Khan, T.I., King, B.P., Frearson, R., Kesteven, P., Wood, P., Daly, A.K., and Wynne, H. (2004). Contribution of age, body size, and CYP2C9 genotype to anticoagulant response to warfarin. Clin. Pharmacol. Ther. *75*, 204–212.

Kawajiri, K., and Fujii-Kuriyama, Y. (2007). Cytochrome P450 gene regulation and physiological functions mediated by the aryl hydrocarbon receptor. Arch. Biochem. Biophys. *464*, 207–212.

Keinan, A., Mullikin, J.C., Patterson, N., and Reich, D. (2007). Measurement of the human allele frequency spectrum demonstrates greater genetic drift in East Asians than in Europeans. Nat. Genet. *39*, 1251–1255.

Kelley, J.L., Madeoy, J., Calhoun, J.C., Swanson, W., and Akey, J.M. (2006). Genomic signatures of positive selection in humans and the limits of outlier approaches. Genome Res. *16*, 980–989.

Khlifi, R., Messaoud, O., Rebai, A., and Hamza-Chaffai, A. (2013). Polymorphisms in the human cytochrome P450 and arylamine N-acetyltransferase: susceptibility to head and neck cancers. BioMed Res. Int. *2013*, 582768.

Kilbane, A.J., Silbart, L.K., Manis, M., Beitins, I.Z., and Weber, W.W. (1990). Human N-acetylation genotype determination with urinary caffeine metabolites. Clin. Pharmacol. Ther. *47*, 470–477.

Kim, I.-W., Kim, K.I., Chang, H.-J., Yeon, B., Bang, S.-J., Park, T., Kwon, J.-S., Kim, S., and Oh, J.M. (2012). Ethnic variability in the allelic distribution of pharmacogenes

between Korean and other populations. Pharmacogenet. Genomics *22*, 829–836.

Kim, U., Jorgenson, E., Coon, H., Leppert, M., Risch, N., and Drayna, D. (2003). Positional cloning of the human quantitative trait locus underlying taste sensitivity to phenylthiocarbamide. Science *299*, 1221–1225.

Kimura, M. (1968). Evolutionary rate at the molecular level. Nature *217*, 624–626.

Kimura, R., Fujimoto, A., Tokunaga, K., and Ohashi, J. (2007). A practical genome scan for population-specific strong selective sweeps that have reached fixation. PloS One *2*, e286.

Klein, T.E., Chang, J.T., Cho, M.K., Easton, K.L., Fergerson, R., Hewett, M., Lin, Z., Liu, Y., Liu, S., Oliver, D.E., et al. (2001). Integrating genotype and phenotype information: an overview of the PharmGKB project. Pharmacogenetics Research Network and Knowledge Base. Pharmacogenomics J. *1*, 167–170.

Klein, T.E., Altman, R.B., Eriksson, N., Gage, B.F., Kimmel, S.E., Lee, M.T., Limdi, N.A., Page, D., Roden, D.M., Wagner, M.J., et al. (2009). Estimation of the warfarin dose with clinical and pharmacogenetic data. N Engl J Med *360*, 753–764.

Köhle, C., and Bock, K.W. (2007). Coordinate regulation of Phase I and II xenobiotic metabolisms by the Ah receptor and Nrf2. Biochem. Pharmacol. *73*, 1853–1862.

Krähenbühl-Melcher, A., Schlienger, R., Lampert, M., Haschke, M., Drewe, J., and Krähenbühl, S. (2007). Drug-related problems in hospitals: a review of the recent literature. Drug Saf. Int. J. Med. Toxicol. Drug Exp. *30*, 379–407.

Krasowski, M.D., Yasuda, K., Hagey, L.R., and Schuetz, E.G. (2005). Evolution of the pregnane x receptor: adaptation to cross-species differences in biliary bile salts. Mol. Endocrinol. Baltim. Md *19*, 1720–1739.

Krasowski, M.D., Yasuda, K., Hagey, L.R., and Schuetz, E.G. (2005b). Evolutionary selection across the nuclear hormone receptor superfamily with a focus on the NR1I subfamily (vitamin D, pregnane X, and constitutive androstane receptors). Nucl. Recept. *3*, 2.

Krimsky, S. (2012). The short life of a race drug. The Lancet *379*, 114–115.

Kuehl, P., Zhang, J., Lin, Y., Lamba, J., Assem, M., Schuetz, J., Watkins, P.B., Daly, A., Wrighton, S.A., Hall, S.D., et al. (2001). Sequence diversity in CYP3A promoters and characterization of the genetic basis of polymorphic CYP3A5 expression. Nat. Genet. *27*, 383–391.

Kukreti, R., Bhatnagar, P., B-Rao, C., Gupta, S., Madan, B., Das, C., Guleria, R., Athavale, A.U., Brahmachari, S.K., and Ghosh, B. (2005). Beta(2)-adrenergic receptor polymorphisms and response to salbutamol among Indian asthmatics*. Pharmacogenomics *6*, 399–410.

Kumar, S., Qiu, H., Oezguen, N., Herlyn, H., Halpert, J.R., and Wojnowski, L. (2009). Ligand diversity of human and chimpanzee CYP3A4: activation of human

CYP3A4 by lithocholic acid results from positive selection. Drug Metab. Dispos. Biol. Fate Chem. *37*, 1328–1333.

Kwiatkowski, D.P. (2005). How malaria has affected the human genome and what human genetics can teach us about malaria. Am J Hum Genet *77*, 171–192.

Lagnaoui, R., Moore, N., Fach, J., Longy-Boursier, M., and Bégaud, B. (2000). Adverse drug reactions in a department of systemic diseases-oriented internal medicine: prevalence, incidence, direct costs and avoidability. Eur. J. Clin. Pharmacol. *56*, 181–186.

Lai, C.S., Fisher, S.E., Hurst, J.A., Vargha-Khadem, F., and Monaco, A.P. (2001). A forkhead-domain gene is mutated in a severe speech and language disorder. Nature *413*, 519–523.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Langley, M.R., Booker, J.K., Evans, J.P., McLeod, H.L., and Weck, K.E. (2009). Validation of clinical testing for warfarin sensitivity: comparison of CYP2C9-VKORC1 genotyping assays and warfarin-dosing algorithms. J. Mol. Diagn. JMD *11*, 216–225.

Lao, O., de Gruijter, J.M., van Duijn, K., Navarro, A., and Kayser, M. (2007). Signatures of positive selection in genes associated with human skin pigmentation as revealed from analyses of single nucleotide polymorphisms. Ann. Hum. Genet. *71*, 354–369.

Lao, O., Lu, T.T., Nothnagel, M., Junge, O., Freitag-Wolf, S., Caliebe, A., Balascakova, M., Bertranpetit, J., Bindoff, L.A., Comas, D., et al. (2008). Correlation between genetic and geographic structure in Europe. Curr. Biol. CB *18*, 1241–1248.

Lappalainen, T., Salmela, E., Andersen, P.M., Dahlman-Wright, K., Sistonen, P., Savontaus, M.-L., Schreiber, S., Lahermo, P., and Kere, J. (2010). Genomic landscape of positive natural selection in Northern European populations. Eur. J. Hum. Genet. EJHG *18*, 471–478.

Larrey, D., Pessayre, D., and Benhamou, J.P. (1985). [Genetic polymorphism of the hepatic metabolism of drugs]. Gastroentérologie Clin. Biol. *9*, 522–531.

Laurent, R., and Chaix, R. (2012). MHC-dependent mate choice in humans: why genomic patterns from the HapMap European American dataset support the hypothesis. BioEssays News Rev. Mol. Cell. Dev. Biol. *34*, 267–271.

Lazarou, J., Pomeranz, B.H., and Corey, P.N. (1998). Incidence of adverse drug reactions in hospitalized patients: a meta-analysis of prospective studies. JAMA J. Am. Med. Assoc. *279*, 1200–1205.

Leape, L.L., Brennan, T.A., Laird, N., Lawthers, A.G., Localio, A.R., Barnes, B.A., Hebert, L., Newhouse, J.P., Weiler, P.C., and Hiatt, H. (1991). The nature of adverse events in hospitalized patients. Results of the Harvard Medical Practice Study II. N. Engl. J. Med. *324*, 377–384.

Lee, P.H., and Shatkay, H. (2008). F-SNP: computationally predicted functional SNPs for disease association studies. Nucleic Acids Res. *36*, D820–D824.

Lee, P.H., and Shatkay, H. (2009). An integrative scoring system for ranking SNPs by their potential deleterious effects. Bioinformatics *25*, 1048–1055.

Lee, C.R., Goldstein, J.A., and Pieper, J.A. (2002). Cytochrome P450 2C9 polymorphisms: a comprehensive review of the in-vitro and human data. Pharmacogenetics *12*, 251–263.

Lee, M.T., Chen, C.H., Chou, C.H., Lu, L.S., Chuang, H.P., Chen, Y.T., Saleem, A.N., Wen, M.S., Chen, J.J., Wu, J.Y., et al. (2009). Genetic determinants of warfarin dosing in the Han-Chinese population. Pharmacogenomics *10*, 1905–1913.

Lee, S.-W., Chung, L.S.-C., Huang, H.-H., Chuang, T.-Y., Liou, Y.-H., and Wu, L.S.-H. (2010). NAT2 and CYP2E1 polymorphisms and susceptibility to first-line anti-tuberculosis drug-induced hepatitis. Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis. *14*, 622–626.

Leff, M.A., Fretland, A.J., Doll, M.A., and Hein, D.W. (1999). Novel human N-acetyltransferase 2 alleles that differ in mechanism for slow acetylator phenotype. J. Biol. Chem. *274*, 34519–34522.

Leiro-Fernandez, V., Valverde, D., Vázquez-Gallardo, R., Botana-Rial, M., Constenla, L., Agúndez, J.A., and Fernández-Villar, A. (2011). N-acetyltransferase 2 polymorphisms and risk of anti-tuberculosis drug-induced hepatotoxicity in Caucasians. Int. J. Tuberc. Lung Dis. Off. J. Int. Union Tuberc. Lung Dis. *15*, 1403–1408.

Lev, E.I., Patel, R.T., Guthikonda, S., Lopez, D., Bray, P.F., and Kleiman, N.S. (2007). Genetic polymorphisms of the platelet receptors P2Y(12), P2Y(1) and GP IIIa and response to aspirin and clopidogrel. Thromb. Res. *119*, 355–360.

Levine, M.N., Raskob, G., Beyth, R.J., Kearon, C., and Schulman, S. (2004). Hemorrhagic complications of anticoagulant treatment: the Seventh ACCP Conference on Antithrombotic and Thrombolytic Therapy. Chest *126*, 287S–310S.

Levy, S., Sutton, G., Ng, P.C., Feuk, L., Halpern, A.L., Walenz, B.P., Axelrod, N., Huang, J., Kirkness, E.F., Denisov, G., et al. (2007). The diploid genome sequence of an individual human. PLoS Biol. *5*, e254.

Lewin, R. (1987). Africa: cradle of modern humans. Science *237*, 1292–1295.

Li, H., Gu, S., Cai, X., Speed, W.C., Pakstis, A.J., Golub, E.I., Kidd, J.R., and Kidd, K.K. (2008a). Ethnic related selection for an ADH Class I variant within East Asia. PloS One *3*, e1881.

Li, J., Zhang, L., Zhou, H., Stoneking, M., and Tang, K. (2011). Global patterns of genetic diversity and signals of natural selection for human ADME genes. Hum Mol Genet *20*, 528–540.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008b). Worldwide

human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

Li, J.Z., Absher, D.M., Tang, H., Southwick, A.M., Casto, A.M., Ramachandran, S., Cann, H.M., Barsh, G.S., Feldman, M., Cavalli-Sforza, L.L., et al. (2008c). Worldwide human relationships inferred from genome-wide patterns of variation. Science *319*, 1100–1104.

Li, T., Chang, C.Y., Jin, D.Y., Lin, P.J., Khvorova, A., and Stafford, D.W. (2004). Identification of the gene for vitamin K epoxide reductase. Nature *427*, 541–544.

Limdi, N.A., Arnett, D.K., Goldstein, J.A., Beasley, T.M., McGwin, G., Adler, B.K., and Acton, R.T. (2008a). Influence of CYP2C9 and VKORC1 on warfarin dose, anticoagulation attainment and maintenance among European-Americans and African-Americans. Pharmacogenomics *9*, 511–526.

Limdi, N.A., Beasley, T.M., Crowley, M.R., Goldstein, J.A., Rieder, M.J., Flockhart, D.A., Arnett, D.K., Acton, R.T., and Liu, N. (2008b). VKORC1 polymorphisms, haplotypes and haplotype groups on warfarin dose among African-Americans and European-Americans. Pharmacogenomics *9*, 1445–1458.

Limdi, N.A., Wadelius, M., Cavallari, L., Eriksson, N., Crawford, D.C., Lee, M.T., Chen, C.H., Motsinger-Reif, A., Sagreiya, H., Liu, N., et al. (2010). Warfarin pharmacogenetics: a single VKORC1 polymorphism is predictive of dose across 3 racial groups. Blood *115*, 3827–3834.

LINK, K.P. (1959). The discovery of dicumarol and its sequels. Circulation *19*, 97–107.

Liu, J., Ding, D., Wang, X., Chen, Y., Li, R., Zhang, Y., and Luo, R. (2012a). N-acetyltransferase polymorphism and risk of colorectal adenoma and cancer: a pooled analysis of variations from 59 studies. PloS One *7*, e42797.

Liu, Y., Yang, J., Xu, Q., Xu, B., Gao, L., Zhang, Y., Zhang, Y., Wang, H., Lu, C., Zhao, Y., et al. (2012b). Comparative performance of warfarin pharmacogenetic algorithms in Chinese patients. Thromb. Res. *130*, 435–440.

Loh, M., Chua, D., Yao, Y., Soo, R.A., Garrett, K., Zeps, N., Platell, C., Minamoto, T., Kawakami, K., Iacopetta, B., et al. (2013). Can population differences in chemotherapy outcomes be inferred from differences in pharmacogenetic frequencies? Pharmacogenomics J. *13*, 423–429.

Lordelo, G.S., Miranda-Vilela, A.L., Akimoto, A.K., Alves, P.C.Z., Hiragi, C.O., Nonino, A., Daldegan, M.B., Klautau-Guimarães, M.N., and Grisolia, C.K. (2012). Association between methylene tetrahydrofolate reductase and glutathione S-transferase M1 gene polymorphisms and chronic myeloid leukemia in a Brazilian population. Genet. Mol. Res. GMR *11*, 1013–1026.

Louicharoen, C., Patin, E., Paul, R., Nuchprayoon, I., Witoonpanich, B., Peerapittayamongkol, C., Casademont, I., Sura, T., Laird, N.M., Singhasivanon, P., et al. (2009). Positively selected G6PD-Mahidol mutation reduces Plasmodium vivax density in Southeast Asians. Science *326*, 1546–1549.

Lu, Y., Kang, L., Hu, K., Wang, C., Sun, X., Chen, F., Kidd, J.R., Kidd, K.K., and Li, H. (2012). High diversity and no significant selection signal of human ADH1B gene in Tibet. Investig. Genet. *3*, 23.

Lubitz, S.A., Scott, S.A., Rothlauf, E.B., Agarwal, A., Peter, I., Doheny, D., Van Der Zee, S., Jaremko, M., Yoo, C., Desnick, R.J., et al. (2010). Comparative performance of gene-based warfarin dosing algorithms in a multiethnic population. J. Thromb. Haemost. JTH *8*, 1018–1026.

Luca, F., Bubba, G., Basile, M., Brdicka, R., Michalodimitrakis, E., Rickards, O., Vershubsky, G., Quintana-Murci, L., Kozlov, A.I., and Novelletto, A. (2008). Multiple advantageous amino acid variants in the NAT2 gene in human populations. PloS One *3*, e3136.

Luo, X., Kranzler, H.R., Zuo, L., Wang, S., Schork, N.J., and Gelernter, J. (2006). Diplotype trend regression analysis of the ADH gene cluster and the ALDH2 gene: multiple significant associations with alcohol dependence. Am. J. Hum. Genet. *78*, 973–987.

Luo, X., Kranzler, H.R., Zuo, L., Wang, S., Schork, N.J., and Gelernter, J. (2007). Multiple ADH genes modulate risk for drug dependence in both African- and European-Americans. Hum. Mol. Genet. *16*, 380–390.

Macgregor, S., Lind, P.A., Bucholz, K.K., Hansell, N.K., Madden, P.A.F., Richter, M.M., Montgomery, G.W., Martin, N.G., Heath, A.C., and Whitfield, J.B. (2009). Associations of ADH and ALDH2 gene variation with self report alcohol reactions, consumption and dependence: an integrated analysis. Hum. Mol. Genet. *18*, 580–593.

Magalon, H., Patin, E., Austerlitz, F., Hegay, T., Aldashev, A., Quintana-Murci, L., and Heyer, E. (2008). Population genetic diversity of the NAT2 gene supports a role of acetylation in human adaptation to farming in Central Asia. Eur. J. Hum. Genet. EJHG *16*, 243–251.

Majeed, A., and Aylin, P. (2005). The ageing population of the United Kingdom and cardiovascular disease. BMJ *331*, 1362.

Mallal, S., Nolan, D., Witt, C., Masel, G., Martin, A.M., Moore, C., Sayer, D., Castley, A., Mamotte, C., Maxwell, D., et al. (2002). Association between presence of HLA-B*5701, HLA-DR7, and HLA-DQ3 and hypersensitivity to HIV-1 reverse-transcriptase inhibitor abacavir. Lancet *359*, 727–732.

Man, M., Farmen, M., Dumaual, C., Teng, C.H., Moser, B., Irie, S., Noh, G.J., Njau, R., Close, S., Wise, S., et al. (2010). Genetic variation in metabolizing enzyme and transporter genes: comprehensive assessment in 3 major East Asian subpopulations with comparison to Caucasians and Africans. J. Clin. Pharmacol. *50*, 929–940.

Mancinelli, L.M., Frassetto, L., Floren, L.C., Dressler, D., Carrier, S., Bekersky, I., Benet, L.Z., and Christians, U. (2001). The pharmacokinetics and metabolic disposition of tacrolimus: a comparison across ethnic groups. Clin. Pharmacol. Ther. *69*, 24–31.

Mannesse, C.K., Derkx, F.H., de Ridder, M.A., Man in 't Veld, A.J., and van der Cammen, T.J. (2000). Contribution of adverse drug reactions to hospital admission of older patients. Age Ageing *29*, 35–39.

Marth, G., Schuler, G., Yeh, R., Davenport, R., Agarwala, R., Church, D., Wheelan, S., Baker, J., Ward, M., Kholodov, M., et al. (2003). Sequence variations in the public human genome data reflect a bottlenecked population history. Proc. Natl. Acad. Sci. U. S. A. *100*, 376–381.

Marth, G.T., Yu, F., Indap, A.R., Garimella, K., Gravel, S., Leong, W.F., Tyler-Smith, C., Bainbridge, M., Blackwell, T., Zheng-Bradley, X., et al. (2011). The functional spectrum of low-frequency coding variation. Genome Biol. *12*, R84.

Masimirembwa, C.M., and Hasler, J.A. (1997). Genetic Polymorphism of Drug Metabolising Enzymes in African Populations: Implications for the Use of Neuroleptics and Antidepressants. Brain Res. Bull. *44*, 561–571.

Materson, B.J., Reda, D.J., Cushman, W.C., Massie, B.M., Freis, E.D., Kochar, M.S., Hamburger, R.J., Fye, C., Lakshman, R., and Gottdiener, J. (1993). Single-drug therapy for hypertension in men. A comparison of six antihypertensive agents with placebo. The Department of Veterans Affairs Cooperative Study Group on Antihypertensive Agents. N. Engl. J. Med. *328*, 914–921.

Mathieson, I., and McVean, G. (2012). Differential confounding of rare and common variants in spatially structured populations. Nat. Genet. *44*, 243–246.

McCarthy, L.C., Davies, K.J., and Campbell, D.A. (2002). Pharmacogenetics in diverse ethnic populations – implications for drug discovery and development. Pharmacogenomics *3*, 493–506.

McGuire, M.C., Nogueira, C.P., Bartels, C.F., Lightstone, H., Hajra, A., Van der Spek, A.F., Lockridge, O., and La Du, B.N. (1989). Identification of the structural mutation responsible for the dibucaine-resistant (atypical) variant form of human serum cholinesterase. Proc. Natl. Acad. Sci. U. S. A. *86*, 953–957.

McKee, R.W., Binkley, S.B., MacCorquidale, D.W., Thayer, S.A., and Doisy, E.A. (1939a). The isolation of vitamine K2. J Am Chem Soc 1295.

McKee, R.W., Binkley, S.B., MacCorquodale, D.W., Thayer, S.A., and Doisy, E.A. (1939b). THE ISOLATION OF VITAMINS K1 AND K2. J. Am. Chem. Soc. *61*, 1295–1295.

McLean, A.J., and Le Couteur, D.G. (2004). Aging biology and geriatric clinical pharmacology. Pharmacol. Rev. *56*, 163–184.

Van der Meer, F.J., Briët, E., Vandenbroucke, J.P., Srámek, D.I., Versluijs, M.H., and Rosendaal, F.R. (1997). The role of compliance as a cause of instability in oral anticoagulant therapy. Br. J. Haematol. *98*, 893–900.

Marshall, A. (1997). Genset-Abbott deal heralds pharmacogenomics era. Nat. Biotechnol. *15*, 829–830.

Mega, J.L., Simon, T., Collet, J.-P., Anderson, J.L., Antman, E.M., Bliden, K., Cannon, C.P., Danchin, N., Giusti, B., Gurbel, P., et al. (2010). Reduced-function CYP2C19

genotype and risk of adverse clinical outcomes among patients treated with clopidogrel predominantly for PCI: a meta-analysis. JAMA J. Am. Med. Assoc. *304*, 1821–1830.

Messer, P.W., and Petrov, D.A. (2013). Population genomics of rapid adaptation by soft selective sweeps. Trends Ecol. Evol. *28*, 659–669.

Meyer, U.A., Zanger, U.M., and Schwab, M. (2013). Omics and drug response. Annu. Rev. Pharmacol. Toxicol. *53*, 475–502.

Meyers, J.L., Nyman, E., Loukola, A., Rose, R.J., Kaprio, J., and Dick, D.M. (2013). The association between DRD2/ANKK1 and genetically informed measures of alcohol use and problems. Addict. Biol. *18*, 523–536.

Mitchell, C., Gregersen, N., and Krause, A. (2011). Novel CYP2C9 and VKORC1 gene variants associated with warfarin dosage variability in the South African black population. Pharmacogenomics.

MITCHELL, R.S., and BELL, J.C. (1957). Clinical implications of isoniazid, PAS and streptomycin blood levels in pulmonary tuberculosis. Trans. Am. Clin. Climatol. Assoc. *69*, 98–102; discussion 103–105.

Moore, T.J., Cohen, M.R., and Furberg, C.D. (2007). Serious adverse drug events reported to the Food and Drug Administration, 1998-2005. Arch. Intern. Med. *167*, 1752–1759.

Moreno-Estrada, A., Aparicio-Prat, E., Sikora, M., Engelken, J., Ramírez-Soriano, A., Calafell, F., and Bosch, E. (2010). African signatures of recent positive selection in human FOXI1. BMC Evol. Biol. *10*, 267.

Mortensen, H.M., Froment, A., Lema, G., Bodo, J.-M., Ibrahim, M., Nyambo, T.B., Omar, S.A., and Tishkoff, S.A. (2011). Characterization of genetic variation and natural selection at the arylamine N-acetyltransferase genes in global human populations. Pharmacogenomics *12*, 1545–1558.

Mossallam, G.I., Abdel Hamid, T.M., and Samra, M.A. (2006). Glutathione S-transferase GSTM1 and GSTT1 polymorphisms in adult acute myeloid leukemia; its impact on toxicity and response to chemotherapy. J. Egypt. Natl. Cancer Inst. *18*, 264–273.Mossallam, G.I., Abdel Hamid, T.M., and Samra, M.A. (2006). Glutathione S-transferase GSTM1 and GSTT1 polymorphisms in adult acute myeloid leukemia; its impact on toxicity and response to chemotherapy. J. Egypt. Natl. Cancer Inst. *18*, 264–273.

MOTULSKY, A.G. (1957). Drug reactions enzymes, and biochemical genetics. J. Am. Med. Assoc. *165*, 835–837.

Mueller, R.L., and Scheidt, S. (1994). History of drugs for thrombotic disease. Discovery, development, and directions for the future. Circulation *89*, 432–449.

Muszkat, M. (2007). Interethnic differences in drug response: the contribution of genetic variability in beta adrenergic receptor and cytochrome P4502C9. Clin. Pharmacol. Ther. *82*, 215–218.

Muthusamy, K.A., Lian, L.H., Vairavan, N., Chua, K.H., and Waran, V. (2012). Genetic polymorphisms of EGF 5'-UTR and NAT2 857G/A associated with glioma in a case control study of Malaysian patients. Genet. Mol. Res. GMR *11*, 2939–2945.

Nachman, M.W., and Crowell, S.L. (2000). Estimate of the mutation rate per nucleotide in humans. Genetics *156*, 297–304.

NCBI (2013). National Center for Biotechnology Information, United States National Library of Medicine.NCBI dbSNP build 138 for human.

Nelson, M.R., Wegmann, D., Ehm, M.G., Kessner, D., St Jean, P., Verzilli, C., Shen, J., Tang, Z., Bacanu, S.-A., Fraser, D., et al. (2012). An abundance of rare functional variants in 202 drug target genes sequenced in 14,002 people. Science *337*, 100–104.

Nielsen, R., Hellmann, I., Hubisz, M., Bustamante, C., and Clark, A.G. (2007). Recent and ongoing selection in the human genome. Nat. Rev. Genet. *8*, 857–868.

Novembre, J., Johnson, T., Bryc, K., Kutalik, Z., Boyko, A.R., Auton, A., Indap, A., King, K.S., Bergmann, S., Nelson, M.R., et al. (2008). Genes mirror geography within Europe. Nature *456*, 98–101.

O'Donnell, P.H., and Dolan, M.E. (2009). Cancer pharmacoethnicity: ethnic differences in susceptibility to the effects of chemotherapy. Clin. Cancer Res. Off. J. Am. Assoc. Cancer Res. *15*, 4806–4814.

Odyakov, V.F., Titova, T.F., Matveev, K.I., and Krysin, A.P. (1992). Vitamin k3, synthesis, properties, and analysis (review). Pharm. Chem. J. *26*, 622–635.

OMS (1983). Comité OMS d'experts de la standardisation biologique.

Oscarson, M., Hidestrand, M., Johansson, I., and Ingelman-Sundberg, M. (1997). A combination of mutations in the CYP2D6*17 (CYP2D6Z) allele causes alterations in enzyme function. Mol. Pharmacol. *52*, 1034–1040.

Osier, M.V., Pakstis, A.J., Soodyall, H., Comas, D., Goldman, D., Odunsi, A., Okonofua, F., Parnas, J., Schulz, L.O., Bertranpetit, J., et al. (2002). A global perspective on genetic variation at the ADH genes reveals unusual patterns of linkage disequilibrium and diversity. Am. J. Hum. Genet. *71*, 84–99.

Pamboukian, S.V., Nisar, I., Patel, S., Gu, L., McLeod, M., Costanzo, M.R., and Heroux, A. (2008). Factors associated with non-adherence to therapy with warfarin in a population of chronic heart failure patients. Clin. Cardiol. *31*, 30–34.

Pankratz, N., Beecham, G.W., DeStefano, A.L., Dawson, T.M., Doheny, K.F., Factor, S.A., Hamza, T.H., Hung, A.Y., Hyman, B.T., Ivinson, A.J., et al. (2012). Meta-analysis of Parkinson's disease: identification of a novel locus, RIT2. Ann. Neurol. *71*, 370–384.

Patillon, B., Luisi, P., Blanché, H., Patin, E., Cann, H.M., Génin, E., and Sabbagh, A. (2012). Positive selection in the chromosome 16 VKORC1 genomic region has contributed to the variability of anticoagulant response in humans. PloS One *7*, e53049.

Patin, E., Barreiro, L.B., Sabeti, P.C., Austerlitz, F., Luca, F., Sajantila, A., Behar, D.M., Semino, O., Sakuntabhai, A., Guiso, N., et al. (2006a). Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. Am J Hum Genet *78*, 423–436.

Patin, E., Barreiro, L.B., Sabeti, P.C., Austerlitz, F., Luca, F., Sajantila, A., Behar, D.M., Semino, O., Sakuntabhai, A., Guiso, N., et al. (2006b). Deciphering the ancient and complex evolutionary history of human arylamine N-acetyltransferase genes. Am. J. Hum. Genet. *78*, 423–436.

Pena, S.D.J. (2011). The fallacy of racial pharmacogenomics. Braz. J. Med. Biol. Res. Rev. Bras. Pesqui. Médicas E Biológicas Soc. Bras. Biofísica Al *44*, 268–275.

Peng, Y., Shi, H., Qi, X., Xiao, C., Zhong, H., Ma, R.Z., and Su, B. (2010). The ADH1B Arg47His polymorphism in east Asian populations and expansion of rice domestication in history. BMC Evol. Biol. *10*, 15.

Pennisi, E. (2012). Genomics. ENCODE project writes eulogy for junk DNA. Science *337*, 1159, 1161.

Perera, M.A., Cavallari, L.H., Limdi, N.A., Gamazon, E.R., Konkashbaev, A., Daneshjou, R., Pluzhnikov, A., Crawford, D.C., Wang, J., Liu, N., et al. (2013). Genetic variants associated with warfarin dose in African-American individuals: a genome-wide association study. Lancet *382*, 790–796.

Peters, E.J., and McLeod, H.L. (2008). Ability of whole-genome SNP arrays to capture "must have" pharmacogenomic variants. Pharmacogenomics *9*, 1573–1577.

Petrovic, M., van der Cammen, T., and Onder, G. (2012). Adverse drug reactions in older people: detection and prevention. Drugs Aging *29*, 453–462.

Phillips, A.L., Nigro, O., Macolino, K.A., Scarborough, K.C., Doecke, C.J., Angley, M.T., and Shakib, S. (2014). Hospital admissions caused by adverse drug events: an Australian prospective study. Aust. Health Rev. Publ. Aust. Hosp. Assoc. *38*, 51–57.

Pickrell, J.K., Coop, G., Novembre, J., Kudaravalli, S., Li, J.Z., Absher, D., Srinivasan, B.S., Barsh, G.S., Myers, R.M., Feldman, M.W., et al. (2009). Signals of recent positive selection in a worldwide sample of human populations. Genome Res *19*, 826–837.

Pirmohamed, M. (2006). Warfarin: almost 60 years old and still causing problems. Br. J. Clin. Pharmacol. *62*, 509–511.

Pirmohamed, M., James, S., Meakin, S., Green, C., Scott, A.K., Walley, T.J., Farrar, K., Park, B.K., and Breckenridge, A.M. (2004). Adverse drug reactions as cause of admission to hospital: prospective analysis of 18 820 patients. BMJ *329*, 15–19.

Polimanti, R., Piacentini, S., De Angelis, F., De Stefano, G.F., and Fuciarelli, M. (2011). Human GST loci as markers of evolutionary forces: GSTO1*E155del and GSTO1*E208K polymorphisms may be under natural selection induced by environmental arsenic. Dis. Markers *31*, 231–239.

Polimanti, R., Piacentini, S., Manfellotto, D., and Fuciarelli, M. (2012). Human genetic variation of CYP450 superfamily: analysis of functional diversity in worldwide populations. Pharmacogenomics *13*, 1951–1960.

POLLOCK, B.E. (1955). Clinical experience with warfarin (coumadin) sodium, a new anticoagulant. J. Am. Med. Assoc. *159*, 1094–1097.

Pontes, Z.B., Vincent-Viry, M., Gueguen, R., Galteau, M.M., and Siest, G. (1993). Acetylation phenotypes and biological variation in a French Caucasian population. Eur. J. Clin. Chem. Clin. Biochem. J. Forum Eur. Clin. Chem. Soc. *31*, 59–68.

Poon, A.H., Gong, L., Brasch-Andersen, C., Litonjua, A.A., Raby, B.A., Hamid, Q., Laprise, C., Weiss, S.T., Altman, R.B., and Klein, T.E. (2012). Very important pharmacogene summary for VDR. Pharmacogenet. Genomics *22*, 758–763.

Pouyanne, P., Haramburu, F., Imbs, J.L., and Bégaud, B. (2000). Admissions to hospital caused by adverse drug reactions: cross sectional incidence study. French Pharmacovigilance Centres. BMJ *320*, 1036.

Pritchard, J.K., and Di Rienzo, A. (2010). Adaptation - not by sweeps alone. Nat. Rev. Genet. *11*, 665–667.

Pritchard, J.K., Pickrell, J.K., and Coop, G. (2010). The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Curr. Biol. CB *20*, R208–215.

Province, M.A., Goetz, M.P., Brauch, H., Flockhart, D.A., Hebert, J.M., Whaley, R., Suman, V.J., Schroth, W., Winter, S., Zembutsu, H., et al. (2014). CYP2D6 genotype and adjuvant tamoxifen: meta-analysis of heterogeneous study populations. Clin. Pharmacol. Ther. *95*, 216–227.

Prugnolle, F., Manica, A., and Balloux, F. (2005). Geography predicts neutral genetic diversity of human populations. Curr. Biol. CB *15*, R159–160.

Przeworski, M., Hudson, R.R., and Di Rienzo, A. (2000). Adjusting the focus on human variation. Trends Genet. TIG *16*, 296–302.

Purcell, S., Neale, B., Todd-Brown, K., Thomas, L., Ferreira, M.A.R., Bender, D., Maller, J., Sklar, P., de Bakker, P.I.W., Daly, M.J., et al. (2007). PLINK: a tool set for whole-genome association and population-based linkage analyses. Am. J. Hum. Genet. *81*, 559–575.

Qiu, H., Taudien, S., Herlyn, H., Schmitz, J., Zhou, Y., Chen, G., Roberto, R., Rocchi, M., Platzer, M., and Wojnowski, L. (2008). CYP3 phylogenomics: evidence for positive selection of CYP3A4 and CYP3A7. Pharmacogenet. Genomics *18*, 53–66.

Qiu, J., Rønnekleiv, O.K., and Kelly, M.J. (2008). Modulation of hypothalamic neuronal activity through a novel G-protein-coupled estrogen membrane receptor. Steroids *73*, 985–991.

Quintana-Murci, L., and Clark, A.G. (2013). Population genetic tools for dissecting innate immunity in humans. Nat. Rev. Immunol. *13*, 280–293.

Raj, T., Kuchroo, M., Replogle, J.M., Raychaudhuri, S., Stranger, B.E., and De Jager, P.L. (2013). Common risk alleles for inflammatory diseases are targets of recent positive selection. Am. J. Hum. Genet. *92*, 517–529.

Ralph, P., and Coop, G. (2010). Parallel adaptation: one or many waves of advance of an advantageous allele? Genetics *186*, 647–668.

Ramagopalan, S.V., Heger, A., Berlanga, A.J., Maugeri, N.J., Lincoln, M.R., Burrell, A., Handunnetthi, L., Handel, A.E., Disanto, G., Orton, S.-M., et al. (2010). A ChIP-seq defined genome-wide map of vitamin D receptor binding: associations with disease and evolution. Genome Res. *20*, 1352–1360.

Ramos, E., Doumatey, A., Elkahloun, A.G., Shriner, D., Huang, H., Chen, G., Zhou, J., McLeod, H., Adeyemo, A., and Rotimi, C.N. (2013). Pharmacogenomics, ancestry and clinical decision making for global populations. Pharmacogenomics J.

Ramsey, L.B., Bruun, G.H., Yang, W., Treviño, L.R., Vattathil, S., Scheet, P., Cheng, C., Rosner, G.L., Giacomini, K.M., Fan, Y., et al. (2012). Rare versus common variants in pharmacogenetics: SLCO1B1 variation and methotrexate disposition. Genome Res. *22*, 1–8.

Ranciaro, A., Campbell, M.C., Hirbo, J.B., Ko, W.-Y., Froment, A., Anagnostou, P., Kotze, M.J., Ibrahim, M., Nyambo, T., Omar, S.A., et al. (2014). Genetic origins of lactase persistence and the spread of pastoralism in Africa. Am. J. Hum. Genet. *94*, 496–510.

Redwood, M., Taylor, C., Bain, B.J., and Matthews, J.H. (1991). The association of age with dosage requirement for warfarin. Age Ageing *20*, 217–220.

Relling, M.V., Gardner, E.E., Sandborn, W.J., Schmiegelow, K., Pui, C.-H., Yee, S.W., Stein, C.M., Carrillo, M., Evans, W.E., Klein, T.E., et al. (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing. Clin. Pharmacol. Ther. *89*, 387–391.

Relling, M.V., Gardner, E.E., Sandborn, W.J., Schmiegelow, K., Pui, C.-H., Yee, S.W., Stein, C.M., Carrillo, M., Evans, W.E., Hicks, J.K., et al. (2013). Clinical pharmacogenetics implementation consortium guidelines for thiopurine methyltransferase genotype and thiopurine dosing: 2013 update. Clin. Pharmacol. Ther. *93*, 324–325.

Rieder, M.J., Reiner, A.P., Gage, B.F., Nickerson, D.A., Eby, C.S., McLeod, H.L., Blough, D.K., Thummel, K.E., Veenstra, D.L., and Rettie, A.E. (2005). Effect of VKORC1 haplotypes on transcriptional regulation and warfarin dose. N Engl J Med *352*, 2285–2293.

Roderick, L.M. (1929). A problem in the coagulation of blood: "sweet clover disease of cattle". J. Am. Vet. Med. Assoc. 314–325.

Rosenberg, N.A., Pritchard, J.K., Weber, J.L., Cann, H.M., Kidd, K.K., Zhivotovsky, L.A., and Feldman, M.W. (2002). Genetic structure of human populations. Science *298*, 2381–2385.

Rosenberg, N.A., Mahajan, S., Ramachandran, S., Zhao, C., Pritchard, J.K., and Feldman, M.W. (2005). Clines, clusters, and the effect of study design on the inference of human population structure. PLoS Genet. *1*, e70.

Ross, K.A., Bigham, A.W., Edwards, M., Gozdzik, A., Suarez-Kurtz, G., and Parra, E.J. (2010). Worldwide allele frequency distribution of four polymorphisms associated with warfarin dose requirements. J Hum Genet *55*, 582–589.

Rost, S., Fregin, A., Ivaskevicius, V., Conzelmann, E., Hortnagel, K., Pelz, H.J., Lappegard, K., Seifried, E., Scharrer, I., Tuddenham, E.G., et al. (2004). Mutations in VKORC1 cause warfarin resistance and multiple coagulation factor deficiency type 2. Nature *427*, 537–541.

Roughead, E.E., Gilbert, A.L., Primrose, J.G., and Sansom, L.N. (1998). Drug-related hospital admissions: a review of Australian studies published 1988-1996. Med. J. Aust. *168*, 405–408.

Ruiz, J.D., Martínez, C., Anderson, K., Gross, M., Lang, N.P., García-Martín, E., and Agúndez, J.A.G. (2012). The differential effect of NAT2 variant alleles permits refinement in phenotype inference and identifies a very slow acetylation genotype. PloS One *7*, e44629.

Rutledge, D.R., Steinberg, J., and Cardozo, L. (1989). Racial differences in drug response: isoproterenol effects on heart rate following intravenous metoprolol. Clin. Pharmacol. Ther. *45*, 380–386.

Ruwende, C., Khoo, S.C., Snow, R.W., Yates, S.N., Kwiatkowski, D., Gupta, S., Warn, P., Allsopp, C.E., Gilbert, S.C., and Peschu, N. (1995). Natural selection of hemi- and heterozygotes for G6PD deficiency in Africa by resistance to severe malaria. Nature *376*, 246–249.

Sabbagh, A., Langaney, A., Darlu, P., Gérard, N., Krishnamoorthy, R., and Poloni, E.S. (2008). Worldwide distribution of NAT2 diversity: implications for NAT2 evolutionary history. BMC Genet. *9*, 21.

Sabbagh, A., Darlu, P., Crouau-Roy, B., and Poloni, E.S. (2011). Arylamine N-acetyltransferase 2 (NAT2) genetic diversity and traditional subsistence: a worldwide population survey. PloS One *6*, e18507.

Sabeti, P.C., Reich, D.E., Higgins, J.M., Levine, H.Z., Richter, D.J., Schaffner, S.F., Gabriel, S.B., Platko, J.V., Patterson, N.J., McDonald, G.J., et al. (2002). Detecting recent positive selection in the human genome from haplotype structure. Nature *419*, 832–837.

Sabeti, P.C., Schaffner, S.F., Fry, B., Lohmueller, J., Varilly, P., Shamovsky, O., Palma, A., Mikkelsen, T.S., Altshuler, D., and Lander, E.S. (2006). Positive natural selection in the human lineage. Science *312*, 1614–1620.

Sabeti, P.C., Varilly, P., Fry, B., Lohmueller, J., Hostetter, E., Cotsapas, C., Xie, X., Byrne, E.H., McCarroll, S.A., Gaudet, R., et al. (2007). Genome-wide detection and characterization of positive selection in human populations. Nature *449*, 913–918.

Sachidanandam, R., Weissman, D., Schmidt, S.C., Kakol, J.M., Stein, L.D., Marth, G., Sherry, S., Mullikin, J.C., Mortimore, B.J., Willey, D.L., et al. (2001). A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. Nature *409*, 928–933.

Sajantila, A., Salem, A.H., Savolainen, P., Bauer, K., Gierig, C., and Pääbo, S. (1996). Paternal and maternal DNA lineages reveal a bottleneck in the founding of the Finnish population. Proc. Natl. Acad. Sci. U. S. A. *93*, 12035–12039.

Salari, K., Watkins, H., and Ashley, E.A. (2012). Personalized medicine: hope or hype? Eur. Heart J. *33*, 1564–1570.

Sanderson, S., Salanti, G., and Higgins, J. (2007). Joint effects of the N-acetyltransferase 1 and 2 (NAT1 and NAT2) genes and smoking on bladder carcinogenesis: a literature-based systematic HuGE review and evidence synthesis. Am. J. Epidemiol. *166*, 741–751.

Saruwatari, J., Nakagawa, K., Shindo, J., Tajiri, T., Fujieda, M., Yamazaki, H., Kamataki, T., and Ishizaki, T. (2002). A population phenotyping study of three drug-metabolizing enzymes in Kyushu, Japan, with use of the caffeine test. Clin. Pharmacol. Ther. *72*, 200–208.

Schelleman, H., Limdi, N.A., and Kimmel, S.E. (2008). Ethnic differences in warfarin maintenance dose requirement and its relationship with genetics. Pharmacogenomics 9, 1331–1346.

Schenekar, T., Winkler, K.A., Troyer, J.L., and Weiss, S. (2011). Isolation and characterization of the CYP2D6 gene in Felidae with comparison to other mammals. J. Mol. Evol. *72*, 222–231.

Schierup, M.H., Mikkelsen, A.M., and Hein, J. (2001). Recombination, balancing selection and phylogenies in MHC and self-incompatibility genes. Genetics *159*, 1833–1844.

Schirmer, M., Toliat, M.R., Haberl, M., Suk, A., Kamdem, L.K., Klein, K., Brockmöller, J., Nürnberg, P., Zanger, U.M., and Wojnowski, L. (2006). Genetic signature consistent with selection against the CYP3A4*1B allele in non-African populations. Pharmacogenet. Genomics *16*, 59–71.

Schlebusch, C.M., Sjödin, P., Skoglund, P., and Jakobsson, M. (2013). Stronger signal of recent selection for lactase persistence in Maasai than in Europeans. Eur. J. Hum. Genet. EJHG *21*, 550–553.

Schofield, F.W. (1924). The cause of a new disease in cattle stimulating hemorrhagic septicaemia and blackleg. J. Am. Vet. Med. Assoc. 553–575.

Sconce, E.A., Khan, T.I., Wynne, H.A., Avery, P., Monkhouse, L., King, B.P., Wood, P., Kesteven, P., Daly, A.K., and Kamali, F. (2005). The impact of CYP2C9 and VKORC1 genetic polymorphism and patient characteristics upon warfarin dose requirements: proposal for a new dosing regimen. Blood *106*, 2329–2333.

Scott, S.A., Sangkuhl, K., Gardner, E.E., Stein, C.M., Hulot, J.-S., Johnson, J.A., Roden, D.M., Klein, T.E., Shuldiner, A.R., and Clinical Pharmacogenetics Implementation Consortium (2011). Clinical Pharmacogenetics Implementation Consortium guidelines for cytochrome P450-2C19 (CYP2C19) genotype and clopidogrel therapy. Clin. Pharmacol. Ther. *90*, 328–332.

Scott, S.A. (2011). Personalizing medicine with clinical pharmacogenetics. Genet. Med. Off. J. Am. Coll. Med. Genet. 13, 987–995.

SEARCH Collaborative Group, Link, E., Parish, S., Armitage, J., Bowman, L., Heath, S., Matsuda, F., Gut, I., Lathrop, M., and Collins, R. (2008). SLCO1B1 variants and statin-induced myopathy--a genomewide study. N. Engl. J. Med. *359*, 789–799.

Selinski, S., Blaszkewicz, M., Ickstadt, K., Hengstler, J.G., and Golka, K. (2013). Refinement of the prediction of N-acetyltransferase 2 (NAT2) phenotypes with respect to enzyme activity and urinary bladder cancer risk. Arch. Toxicol. *87*, 2129–2139.

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res. *29*, 308–311.

Shin, J., and Cao, D. (2011). Comparison of warfarin pharmacogenetic dosing algorithms in a racially diverse large cohort. Pharmacogenomics *12*, 125–134.

Shuldiner, A.R., O'Connell, J.R., Bliden, K.P., Gandhi, A., Ryan, K., Horenstein, R.B., Damcott, C.M., Pakyz, R., Tantry, U.S., Gibson, Q., et al. (2009). Association of cytochrome P450 2C19 genotype with the antiplatelet effect and clinical efficacy of clopidogrel therapy. JAMA J. Am. Med. Assoc. *302*, 849–857.

Sim, S.C., and Ingelman-Sundberg, M. (2011). Pharmacogenomic biomarkers: new tools in current and future drug therapy. Trends Pharmacol. Sci. *32*, 72–81.

Sim, S.C., Altman, R.B., and Ingelman-Sundberg, M. (2011). Databases in the area of pharmacogenetics. Hum. Mutat. *32*, 526–531.

Sistonen, J., Fuselli, S., Palo, J.U., Chauhan, N., Padh, H., and Sajantila, A. (2009). Pharmacogenetic variation at CYP2C9, CYP2C19, and CYP2D6 at global and microgeographic scales. Pharmacogenet Genomics *19*, 170–179.

Slatkin, M., and Excoffier, L. (2012). Serial founder effects during range expansion: a spatial analog of genetic drift. Genetics *191*, 171–181.

Smith, J.M., and Haigh, J. (1974). The hitch-hiking effect of a favourable gene. Genet. Res. *23*, 23–35.

Snyder, L.H. (1931). INHERITED TASTE DEFICIENCY. Science *74*, 151–152.

Spear, B.B., Heath-Chiozzi, M., and Huff, J. (2001). Clinical application of pharmacogenetics. Trends Mol. Med. *7*, 201–204.

Stenflo, J., Fernlund, P., Egan, W., and Roepstorff, P. (1974). Vitamin K dependent modifications of glutamic acid residues in prothrombin. Proc. Natl. Acad. Sci. U. S. A. *71*, 2730–2733.

Stephens, J.C., Schneider, J.A., Tanguay, D.A., Choi, J., Acharya, T., Stanley, S.E., Jiang, R., Messer, C.J., Chew, A., Han, J.H., et al. (2001). Haplotype variation and linkage disequilibrium in 313 human genes. Science *293*, 489–493.

Strachan, D.P. (2000). Family size, infection and atopy: the first decade of the "hygiene hypothesis."Thorax *55 Suppl 1*, S2–10.

Stuart, P.E., Nair, R.P., Ellinghaus, E., Ding, J., Tejasvi, T., Gudjonsson, J.E., Li, Y., Weidinger, S., Eberlein, B., Gieger, C., et al. (2010). Genome-wide association analysis identifies three psoriasis susceptibility loci. Nat. Genet. *42*, 1000–1004.

Suarez-Kurtz, G., Pena, S.D.J., and Hutz, M.H. (2012a). Application of the F(ST) statistics to explore pharmacogenomic diversity in the Brazilian population. Pharmacogenomics *13*, 771–777.

Suarez-Kurtz, G., Pena, S.D.J., Struchiner, C.J., and Hutz, M.H. (2012b). Pharmacogenomic Diversity among Brazilians: Influence of Ancestry, Self-Reported Color, and Geographical Origin. Front. Pharmacol. *3*, 191.

Suarez-Kurtz, G., Genro, J.P., de Moraes, M.O., Ojopi, E.B., Pena, S.D.J., Perini, J.A., Ribeiro-dos-Santos, A., Romano-Silva, M.A., Santana, I., and Struchiner, C.J. (2012c). Global pharmacogenomics: Impact of population diversity on the distribution of polymorphisms in the CYP2C cluster among Brazilians. Pharmacogenomics J. *12*, 267–276.

Suarez-Kurtz, G.M., Pena, S.D.J.M., Struchiner, C.J.M., and Hutz, M.H.P. (2012d). Pharmacogenomic diversity among Brazilians: influence of ancestry, self-reported color, and geographical origin. Pharmacogenetics Pharmacogenomics *3*, 191.

Sung, C., Lee, P.L., Tan, L.L., and Toh, D.S.L. (2011). Pharmacogenetic risk for adverse reactions to irinotecan in the major ethnic populations of Singapore: regulatory evaluation by the health sciences authority. Drug Saf. Int. J. Med. Toxicol. Drug Exp. *34*, 1167–1175.

Tajima, F. (1989a). Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics *123*, 585–595.

Tajima, F. (1989b). The Effect of Change in Population Size on DNA Polymorphism. Genetics *123*, 597.

Takahashi, H., Wilkinson, G.R., Nutescu, E.A., Morita, T., Ritchie, M.D., Scordo, M.G., Pengo, V., Barban, M., Padrini, R., Ieiri, I., et al. (2006). Different contributions of polymorphisms in VKORC1 and CYP2C9 to intra- and inter-population differences in maintenance dose of warfarin in Japanese, Caucasians and African-Americans. Pharmacogenet Genomics *16*, 101–110.

Takeuchi, F., McGinnis, R., Bourgeois, S., Barnes, C., Eriksson, N., Soranzo, N., Whittaker, P., Ranganath, V., Kumanduri, V., McLaren, W., et al. (2009). A genome-wide association study confirms VKORC1, CYP2C9, and CYP4F2 as principal genetic determinants of warfarin dose. PLoS Genet *5*, e1000433.

Tang, B.K., Kadar, D., and Kalow, W. (1987). An alternative test for acetylator phenotyping with caffeine. Clin. Pharmacol. Ther. *42*, 509–513.

Tang, B.K., Kadar, D., Qian, L., Iriah, J., Yip, J., and Kalow, W. (1991). Caffeine as a metabolic probe: validation of its use for acetylator phenotyping. Clin. Pharmacol. Ther. *49*, 648–657.

Tang, K., Thornton, K.R., and Stoneking, M. (2007). A new approach for using genome scans to detect recent positive selection in the human genome. PLoS Biol. *5*, e171.

Taylor, A.L., Ziesche, S., Yancy, C., Carson, P., D'Agostino, R., Jr, Ferdinand, K., Taylor, M., Adams, K., Sabolinski, M., Worcel, M., et al. (2004). Combination of isosorbide dinitrate and hydralazine in blacks with heart failure. N. Engl. J. Med. *351*, 2049–2057.

Teichert, M., Eijgelsheim, M., Rivadeneira, F., Uitterlinden, A.G., van Schaik, R.H., Hofman, A., De Smet, P.A., van Gelder, T., Visser, L.E., and Stricker, B.H. (2009). A genome-wide association study of acenocoumarol maintenance dosage. Hum Mol Genet *18*, 3758–3768.

Teixeira, R.L. de F., Morato, R.G., Cabello, P.H., Muniz, L.M.K., Moreira, A. da S.R., Kritski, A.L., Mello, F.C.Q., Suffys, P.N., Miranda, A.B. de, and Santos, A.R. (2011). Genetic polymorphisms of NAT2, CYP2E1 and GST enzymes and the occurrence of antituberculosis drug-induced hepatitis in Brazilian TB patients. Mem. Inst. Oswaldo Cruz *106*, 716–724.

Tennessen, J.A., Bigham, A.W., O'Connor, T.D., Fu, W., Kenny, E.E., Gravel, S., McGee, S., Do, R., Liu, X., Jun, G., et al. (2012). Evolution and functional impact of rare coding variation from deep sequencing of human exomes. Science *337*, 64–69.

Teo, Y.-Y., Sim, X., Ong, R.T.H., Tan, A.K.S., Chen, J., Tantoso, E., Small, K.S., Ku, C.-S., Lee, E.J.D., Seielstad, M., et al. (2009). Singapore Genome Variation Project: a haplotype map of three Southeast Asian populations. Genome Res. *19*, 2154–2162.

Teo, Y.-Y., Small, K.S., and Kwiatkowski, D.P. (2010). Methodological challenges of genome-wide association analysis in Africa. Nat. Rev. Genet. *11*, 149–160.

Thomas, R.S., Penn, S.G., Holden, K., Bradfield, C.A., and Rank, D.R. (2002). Sequence variation and phylogenetic history of the mouse Ahr gene. Pharmacogenetics *12*, 151–163.

Thompson, E.E., Kuttab-Boulos, H., Witonsky, D., Yang, L., Roe, B.A., and Di Rienzo, A. (2004). CYP3A Variation and the Evolution of Salt-Sensitivity Variants. Am. J. Hum. Genet. *75*, 1059.

Thomson, R., Pritchard, J.K., Shen, P., Oefner, P.J., and Feldman, M.W. (2000). Recent common ancestry of human Y chromosomes: evidence from DNA sequence data. Proc. Natl. Acad. Sci. U. S. A. *97*, 7360–7365.

Tishkoff, S.A., Varkonyi, R., Cahinhinan, N., Abbes, S., Argyropoulos, G., Destro-Bisol, G., Drousiotou, A., Dangerfield, B., Lefranc, G., Loiselet, J., et al. (2001). Haplotype diversity and linkage disequilibrium at human G6PD: recent origin of alleles that confer malarial resistance. Science *293*, 455–462.

Tishkoff, S.A., Reed, F.A., Ranciaro, A., Voight, B.F., Babbitt, C.C., Silverman, J.S., Powell, K., Mortensen, H.M., Hirbo, J.B., Osman, M., et al. (2007). Convergent adaptation of human lactase persistence in Africa and Europe. Nat. Genet. *39*, 31–40.

Tsukamoto, Y., Ichise, H., Kakuda, H., and Yamaguchi, M. (2000). Intake of fermented soybean (natto) increases circulating vitamin K2 (menaquinone-7) and gamma-carboxylated osteocalcin concentration in normal individuals. J. Bone Miner. Metab. *18*, 216–222.

Uetrecht, J., and Naisbitt, D.J. (2013). Idiosyncratic adverse drug reactions: current concepts. Pharmacol. Rev. *65*, 779–808.

Vargha-Khadem, F., Watkins, K., Alcock, K., Fletcher, P., and Passingham, R. (1995). Praxic and nonverbal cognitive deficits in a large family with a genetically transmitted speech and language disorder. Proc. Natl. Acad. Sci. U. S. A. *92*, 930–933.

Vasseur, E., and Quintana-Murci, L. (2013). The impact of natural selection on health and disease: uses of the population genetics approach in humans. Evol. Appl. *6*, 596–607.

Venter, C.P., and Joubert, P.H. (1984). Ethnic differences in response to beta 1-adrenoceptor blockade by propranolol. J. Cardiovasc. Pharmacol. *6*, 361–364.

Venter, J.C., Adams, M.D., Myers, E.W., Li, P.W., Mural, R.J., Sutton, G.G., Smith, H.O., Yandell, M., Evans, C.A., Holt, R.A., et al. (2001). The sequence of the human genome. Science *291*, 1304–1351.

Vermeer, C., Shearer, M.J., Zittermann, A., Bolton-Smith, C., Szulc, P., Hodges, S., Walter, P., Rambeck, W., Stöcklin, E., and Weber, P. (2004). Beyond deficiency: potential benefits of increased intakes of vitamin K for bone and vascular health. Eur. J. Nutr. *43*, 325–335.

Vesell, E.S. (1978). Twin studies in pharmacogenetics. Hum. Genet. Suppl. 19–30.

Vogel, F. (1959). Moderne Probleme der Humangenetik. Ergebn Inn Med Kinderheilkd 52–125.

Voight, B.F., Kudaravalli, S., Wen, X., and Pritchard, J.K. (2006). A map of recent positive selection in the human genome. PLoS Biol *4*, e72.

Wadelius, M., Chen, L.Y., Downes, K., Ghori, J., Hunt, S., Eriksson, N., Wallerman, O., Melhus, H., Wadelius, C., Bentley, D., et al. (2005). Common VKORC1 and GGCX polymorphisms associated with warfarin dose. Pharmacogenomics J *5*, 262–270.

Wang, D., Chen, H., Momary, K.M., Cavallari, L.H., Johnson, J.A., and Sadee, W. (2008). Regulatory polymorphism in vitamin K epoxide reductase complex subunit 1 (VKORC1) affects gene expression and warfarin dose requirement. Blood *112*, 1013–1021.

Wang, D., and Sadee, W. (2012). The Making of a CYP3A Biomarker Panel for Guiding Drug Therapy. J. Pers. Med. *2*, 175–191.

Wang, E.T., Kodama, G., Baldi, P., and Moyzis, R.K. (2006). Global landscape of recent inferred Darwinian selection for Homo sapiens. Proc. Natl. Acad. Sci. U. S. A. *103*, 135–140.

Wang, Y., Zhang, Q., Zhang, M., and Wang, C. (2014). NAT2 slow acetylation genotypes contribute to asthma risk among Caucasians: evidence from 946 cases and 1,091 controls. Mol. Biol. Rep. *41*, 1849–1855.

Wardrop, D., and Keeling, D. (2008). The story of the discovery of heparin and warfarin. Br. J. Haematol. *141*, 757–763.

Waterman, A.D., Milligan, P.E., Bayer, L., Banet, G.A., Gatchel, S.K., and Gage, B.F. (2004). Effect of warfarin nonadherence on control of the International Normalized Ratio. Am. J. Health-Syst. Pharm. AJHP Off. J. Am. Soc. Health-Syst. Pharm. *61*, 1258–1264.

WATSON, J.D., and CRICK, F.H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature *171*, 737–738.

Weber, W.W., and Hein, D.W. (1985). N-acetylation pharmacogenetics. Pharmacol. Rev. *37*, 25–79.

Wedekind, C., and Penn, D. (2000). MHC genes, body odours, and odour preferences. Nephrol. Dial. Transplant. Off. Publ. Eur. Dial. Transpl. Assoc. - Eur. Ren. Assoc. *15*, 1269–1271.

Weir, B.S., and Cockerham, C.C. (1984). Estimating F-statistics for the analysis of population structure. Evolution 1358–1370.

Weir, B.S., Cardon, L.R., Anderson, A.D., Nielsen, D.M., and Hill, W.G. (2005). Measures of human population structure show heterogeneity among genomic regions. Genome Res *15*, 1468–1476.

Wells, P.S., Holbrook, A.M., Crowther, N.R., and Hirsh, J. (1994). Interactions of warfarin with drugs and food. Ann. Intern. Med. *121*, 676–683.

Whitlon, D.S., Sadowski, J.A., and Suttie, J.W. (1978). Mechanism of coumarin action: significance of vitamin K epoxide reductase inhibition. Biochemistry (Mosc.) *17*, 1371–1377.

Wilke, R.A., Ramsey, L.B., Johnson, S.G., Maxwell, W.D., McLeod, H.L., Voora, D., Krauss, R.M., Roden, D.M., Feng, Q., Cooper-Dehoff, R.M., et al. (2012). The clinical pharmacogenomics implementation consortium: CPIC guideline for SLCO1B1 and simvastatin-induced myopathy. Clin. Pharmacol. Ther. 92, 112–117.

Williamson, S.H., Hubisz, M.J., Clark, A.G., Payseur, B.A., Bustamante, C.D., and Nielsen, R. (2007). Localizing recent adaptive evolution in the human genome. PLoS Genet. *3*, e90.

Wilson, K. (1984). Sex-related differences in drug disposition in man. Clin. Pharmacokinet. *9*, 189–202.

Wilson, J.F., Weale, M.E., Smith, A.C., Gratrix, F., Fletcher, B., Thomas, M.G., Bradman, N., and Goldstein, D.B. (2001). Population genetic structure of variable drug response. Nat. Genet. *29*, 265–269.

Wood, A.J. (1998). Ethnic differences in drug disposition and response. Ther. Drug Monit. *20*, 525–526.

Wooding, S.P., Watkins, W.S., Bamshad, M.J., Dunn, D.M., Weiss, R.B., and Jorde, L.B. (2002). DNA sequence variation in a 3.7-kb noncoding sequence 5' of the CYP1A2 gene: implications for human population history and natural selection. Am. J. Hum. Genet. *71*, 528–542.

World Health Organization (1969). International Drug Monitoring: The Role of the Hospital.

Wright, S. (1931). Evolution in Mendelian Populations. Genetics *16*, 97–159.

WRIGHT, S. (1951). The genetical structure of populations. Ann. Eugen. *15*, 323–354.

Wu, T.-Y., Jen, M.-H., Bottle, A., Molokhia, M., Aylin, P., Bell, D., and Majeed, A. (2010). Ten-year trends in hospital admissions for adverse drug reactions in England 1999-2009. J. R. Soc. Med. *103*, 239–250.

Wynne, H., Cope, L., Kelly, P., Whittingham, T., Edwards, C., and Kamali, F. (1995). The influence of age, liver size and enantiomer concentrations on warfarin requirements. Br. J. Clin. Pharmacol. *40*, 203–207.

Xie, H.G., Kim, R.B., Wood, A.J., and Stein, C.M. (2001). Molecular basis of ethnic differences in drug disposition and response. Annu. Rev. Pharmacol. Toxicol. *41*, 815–850.

Xue, Y., Zhang, X., Huang, N., Daly, A., Gillson, C.J., Macarthur, D.G., Yngvadottir, B., Nica, A.C., Woodwark, C., Chen, Y., et al. (2009). Population differentiation as an indicator of recent positive selection in humans: an empirical evaluation. Genetics *183*, 1065–1077.

Yasuda, S.U., Zhang, L., and Huang, S.-M. (2008). The role of ethnicity in variability in response to drugs: focus on clinical pharmacology studies. Clin. Pharmacol. Ther. *84*, 417–423.

Yen-Revollo, J.L., Auman, J.T., and McLeod, H.L. (2008). Race does not explain genetic heterogeneity in pharmacogenomic pathways. Pharmacogenomics 9, 1639–1645.

Yi, X., Liang, Y., Huerta-Sanchez, E., Jin, X., Cuo, Z.X.P., Pool, J.E., Xu, X., Jiang, H., Vinckenbosch, N., Korneliussen, T.S., et al. (2010). Sequencing of 50 human exomes reveals adaptation to high altitude. Science *329*, 75–78.

Yuan, H.Y., Chen, J.J., Lee, M.T., Wung, J.C., Chen, Y.F., Charng, M.J., Lu, M.J., Hung, C.R., Wei, C.Y., Chen, C.H., et al. (2005). A novel functional VKORC1 promoter polymorphism is associated with inter-individual and inter-ethnic differences in warfarin sensitivity. Hum Mol Genet *14*, 1745–1751.

Zanger, U.M., and Schwab, M. (2013). Cytochrome P450 enzymes in drug metabolism: regulation of gene expression, enzyme activities, and impact of genetic variation. Pharmacol. Ther. *138*, 103–141.

Zanger, U.M., Raimundo, S., and Eichelbaum, M. (2004). Cytochrome P450 2D6: overview and update on pharmacology, genetics, biochemistry. Naunyn. Schmiedebergs Arch. Pharmacol. *369*, 23–37.

Zendeh-Boodi, Z., and Saadat, M. (2008). Genetic polymorphism of GSTT1 may be under natural selection in a population chronically exposed to natural sour gas. Mol. Biol. Rep. *35*, 673–676.

Zhang, C., Bailey, D.K., Awad, T., Liu, G., Xing, G., Cao, M., Valmeekam, V., Retief, J., Matsuzaki, H., Taub, M., et al. (2006). A whole genome long-range haplotype (WGLRH) test for detecting imprints of positive selection in human populations. Bioinforma. Oxf. Engl. *22*, 2122–2128.

Zhang, J., Webb, D.M., and Podlaha, O. (2002). Accelerated protein evolution and origins of human-specific features: Foxp2 as an example. Genetics *162*, 1825–1835.

Zheng, C.J., Han, L.Y., Xie, B., Liew, C.Y., Ong, S., Cui, J., Zhang, H.L., Tang, Z.Q., Gan, S.H., Jiang, L., et al. (2007). PharmGED: Pharmacogenetic Effect Database. Nucleic Acids Res. *35*, D794–D799.

Zhong, M., Lange, K., Papp, J.C., and Fan, R. (2010). A powerful score test to detect positive selection in genome-wide scans. Eur. J. Hum. Genet. EJHG *18*, 1148–1159.

Zhong, M., Zhang, Y., Lange, K., and Fan, R. (2011). A cross-population extended haplotype-based homozygosity score test to detect positive selection in genome-wide scans. Stat. Interface 51–63.

Zhou, L.-P., Yao, F., Luan, H., Wang, Y.-L., Dong, X.-H., Zhou, W.-W., and Wang, Q.-H. (2013). CYP3A4*1B polymorphism and cancer risk: a HuGE review and meta-analysis. Tumour Biol. J. Int. Soc. Oncodevelopmental Biol. Med. *34*, 649–660.

Zmuda, J.M., Cauley, J.A., and Ferrell, R.E. (2000). Molecular epidemiology of vitamin D receptor gene variants. Epidemiol. Rev. *22*, 203–217.

Zuccolo, L., Fitz-Simon, N., Gray, R., Ring, S.M., Sayal, K., Smith, G.D., and Lewis, S.J. (2009). A non-synonymous variant in ADH1B is strongly associated with prenatal alcohol use in a European sample of pregnant women. Hum. Mol. Genet. *18*, 4457–4466.

Annexes

Annexe 1

Tables et Figures supplémentaires de l'article 1

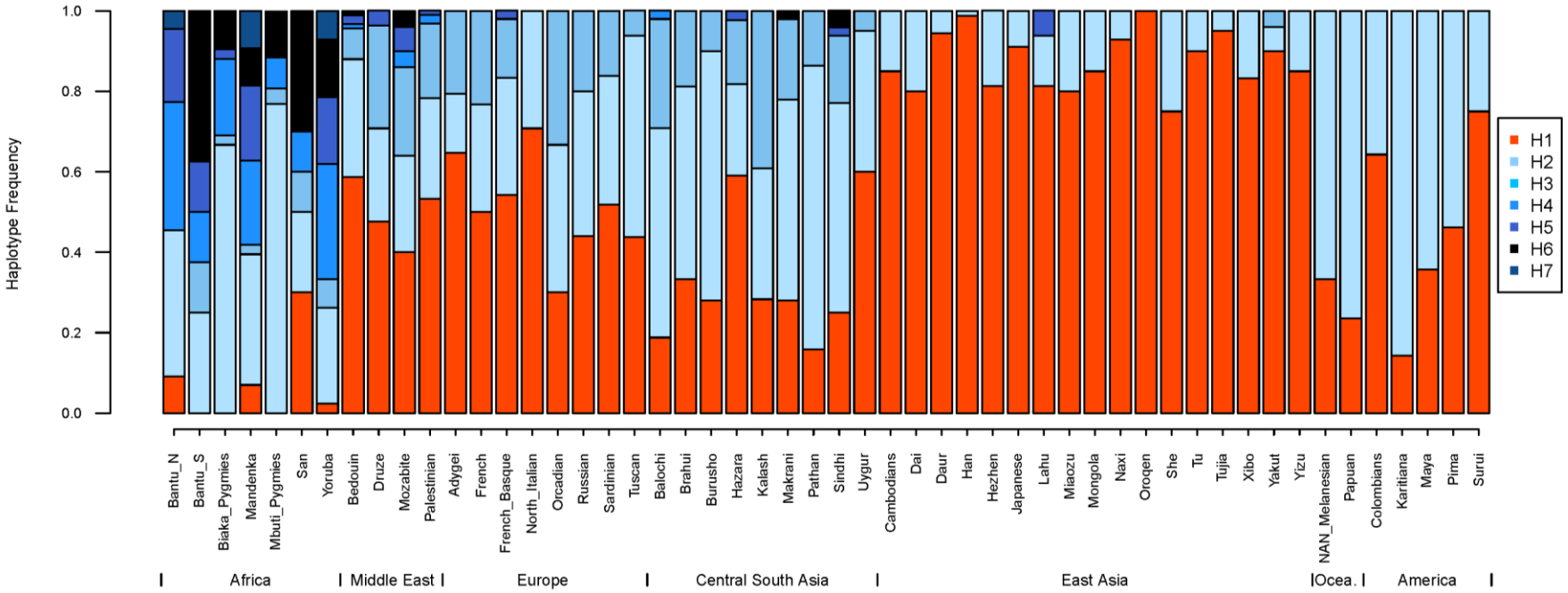


Figure S1. Distribution of *VKORC1* haplotypes in the 52 HGDP-CEPH samples. The haplotype carrying the -1639A allele (H1) is represented in red and the ancestral haplotype (H6) in black.

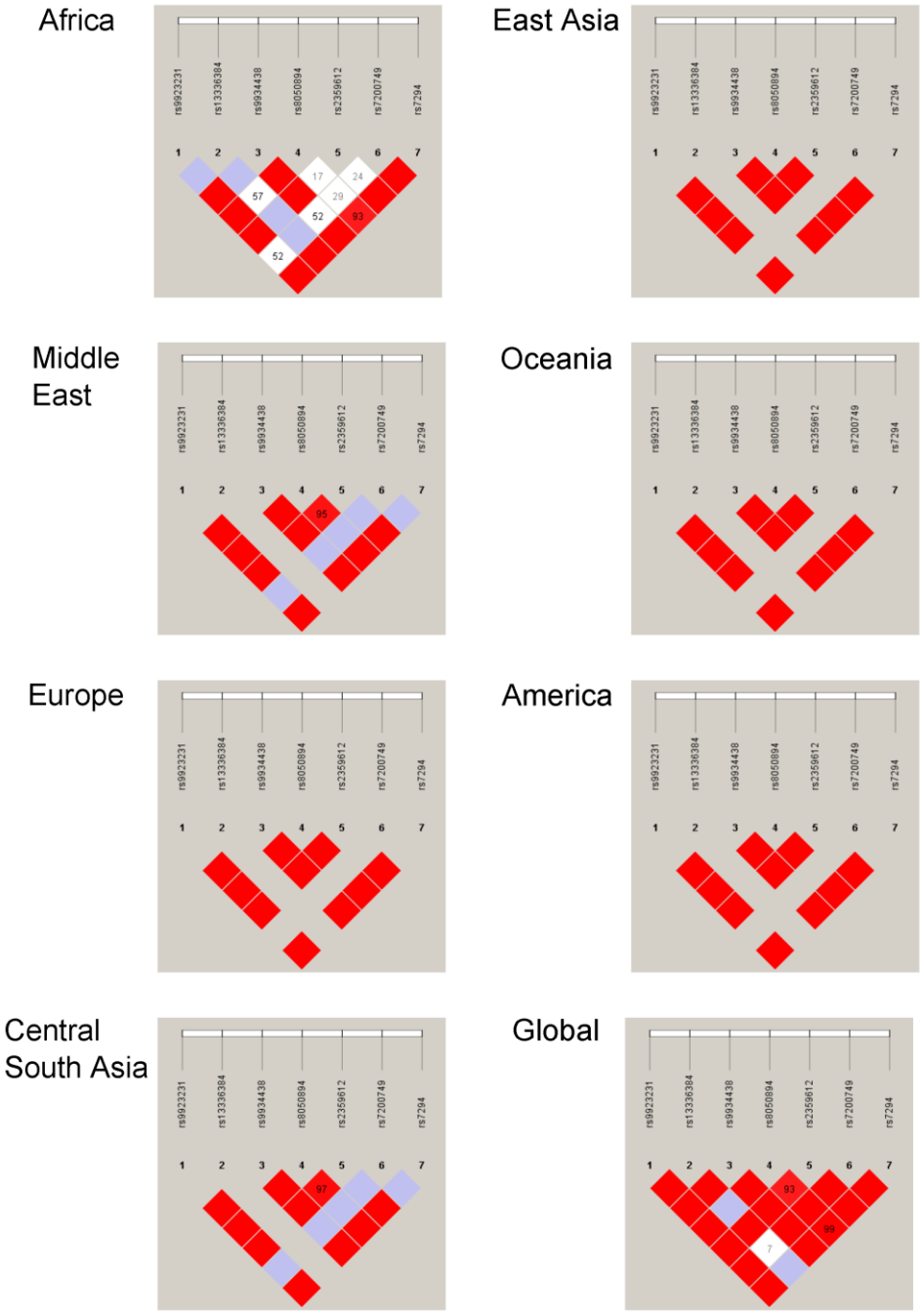


Figure S2. Pairwise LD between the seven *VKORC1* SNPs at the regional and global level. Red squares indicate statistically significant (logarithm of odds > 2) LD between the pair of SNPs, as measured by the *D*' statistic (Lewontin, 1964) with the Haploview software (Barrett et a., 2005) ; darker colors of red indicate higher values of *D*', up to a maximum of 1. White squares indicate pairwise *D*' values of < 1 with no statistically significant evidence of LD. Blue squares indicate pairwise *D*' values of 1 but without statistical significance.

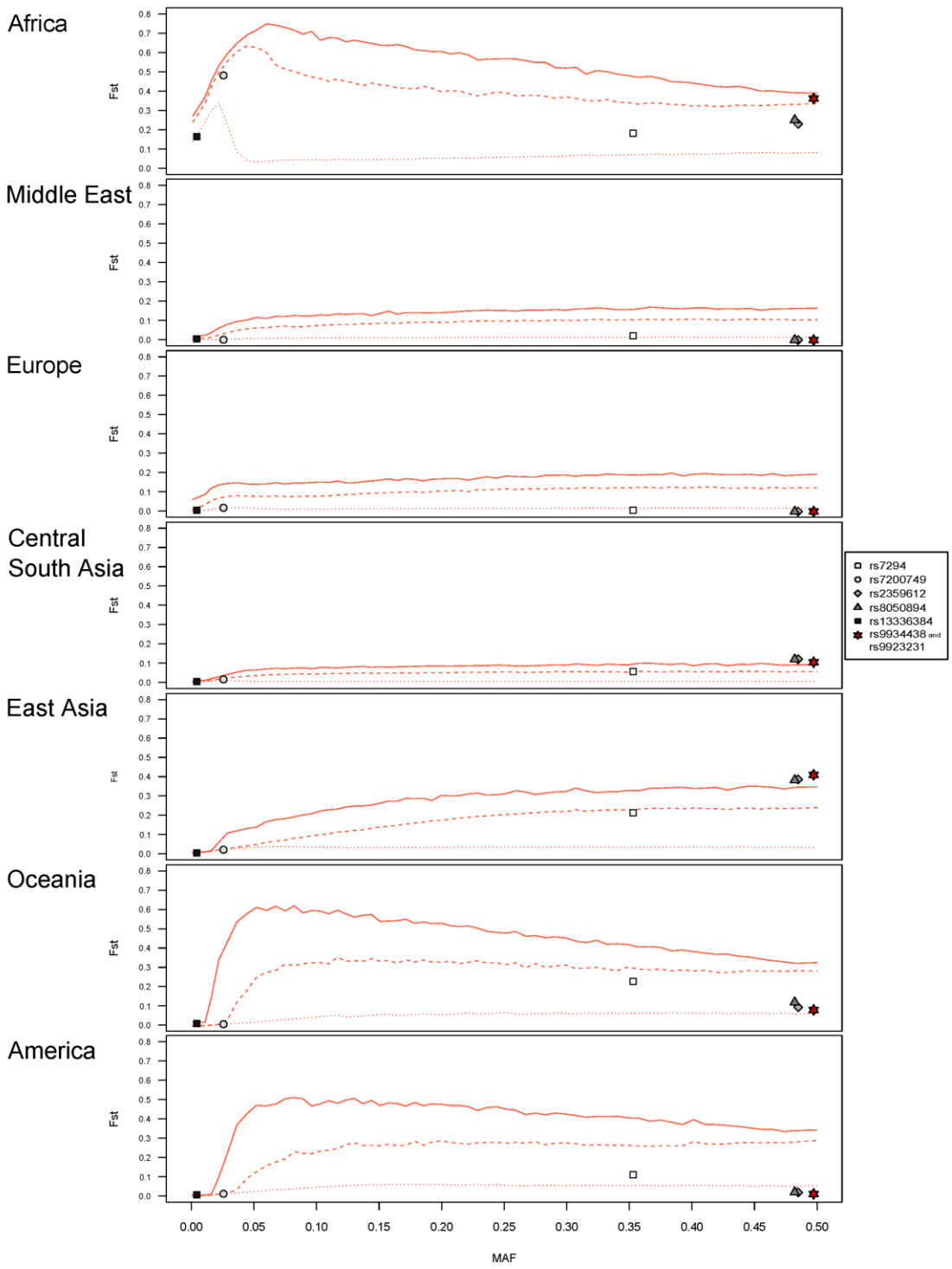


Figure S3. Genome-wide empirical distributions of inter-regional F_{ST} values against MAF in the seven geographic regions. Empirical distributions of F_{ST} were constructed by calculating an F_{ST} value for 644,413 SNPs having a MAF ≥ 0.001 at the global level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against their global MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.

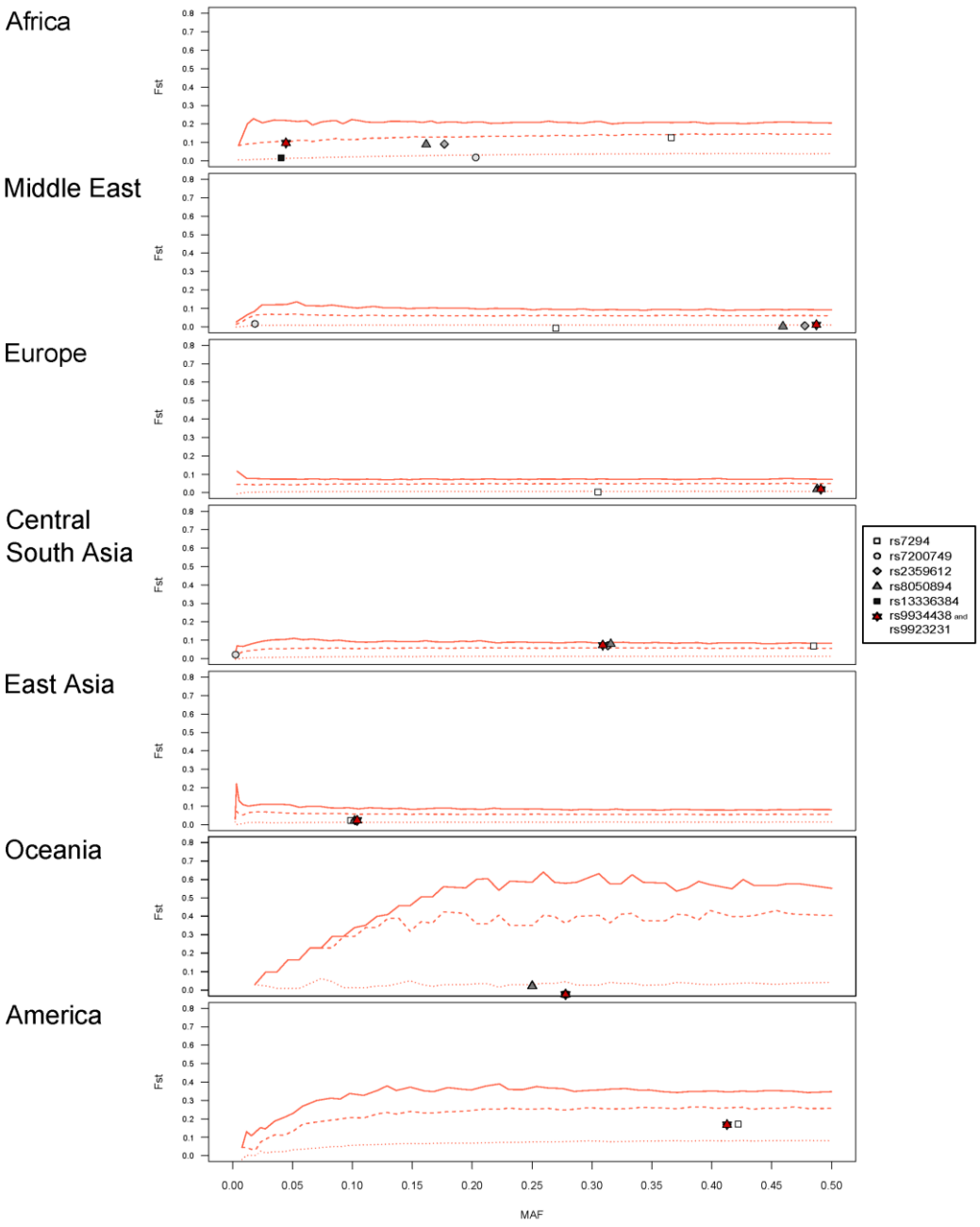


Figure S4. Genome-wide empirical distributions of intra-regional F_{ST} values against MAF in the seven geographic regions. Empirical distributions of F_{ST} were constructed by calculating an F_{ST} value for all SNPs having a MAF ≥ 0.001 at the intra-regional level. Individual values of F_{ST} calculated for each of the seven *VKORC1* SNPs are plotted against the regional MAF. The functional rs9923231 SNP is shown in red. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively.

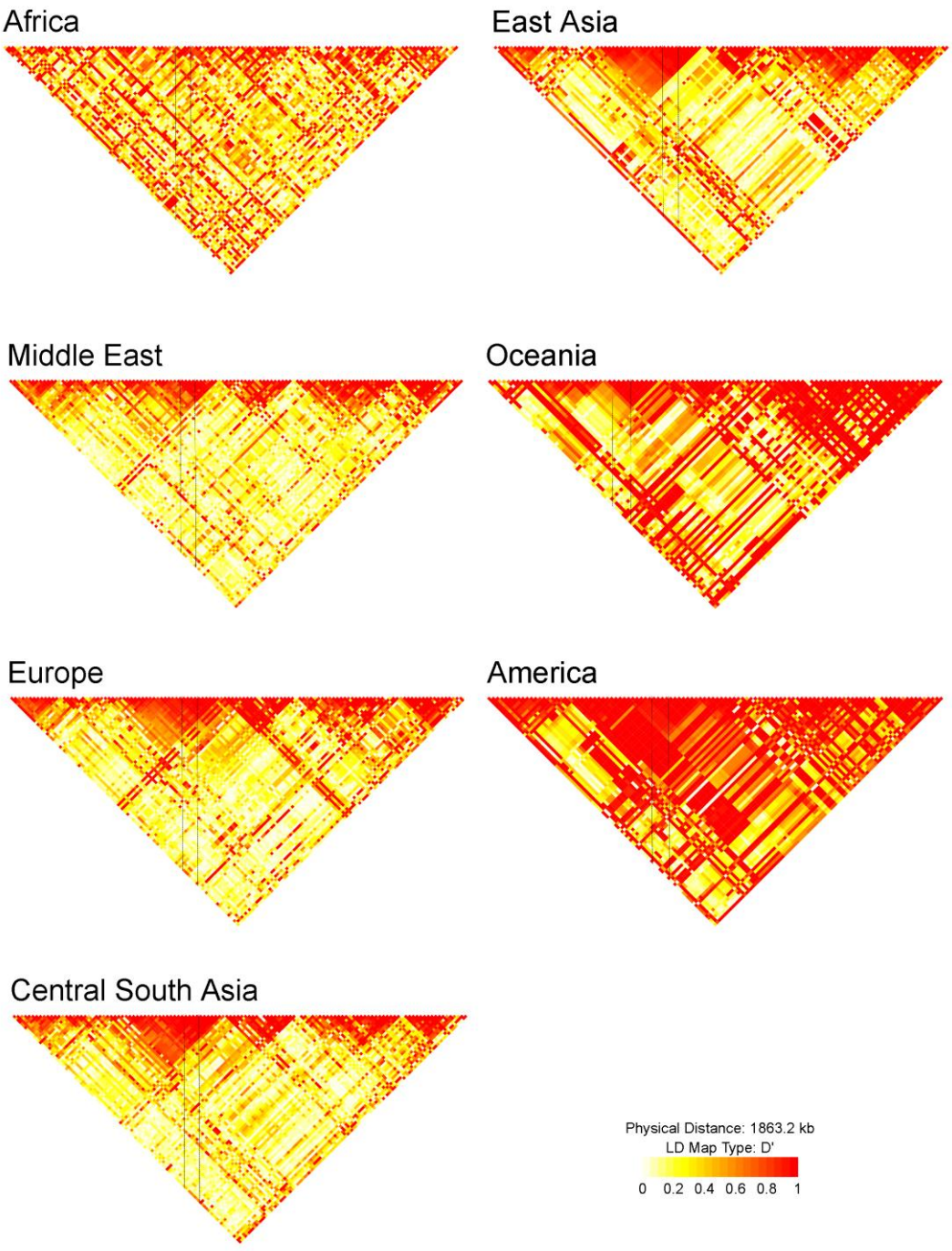


Figure S5. LD patterns over a 2 Mb region centered on *VKORC1* in the seven geographic regions. Pairwise LD, depicted as *D*', is shown for SNPs with a MAF ≥ 0.05 at the global level. *D*' values are displayed in different colors from yellow to red for *D*' = 0 to *D*' = 1, respectively. The plot was produced using the snp.plotter R package (Luna et al., 2007). The vertical dashed lines delineate *VKORC1* gene position.

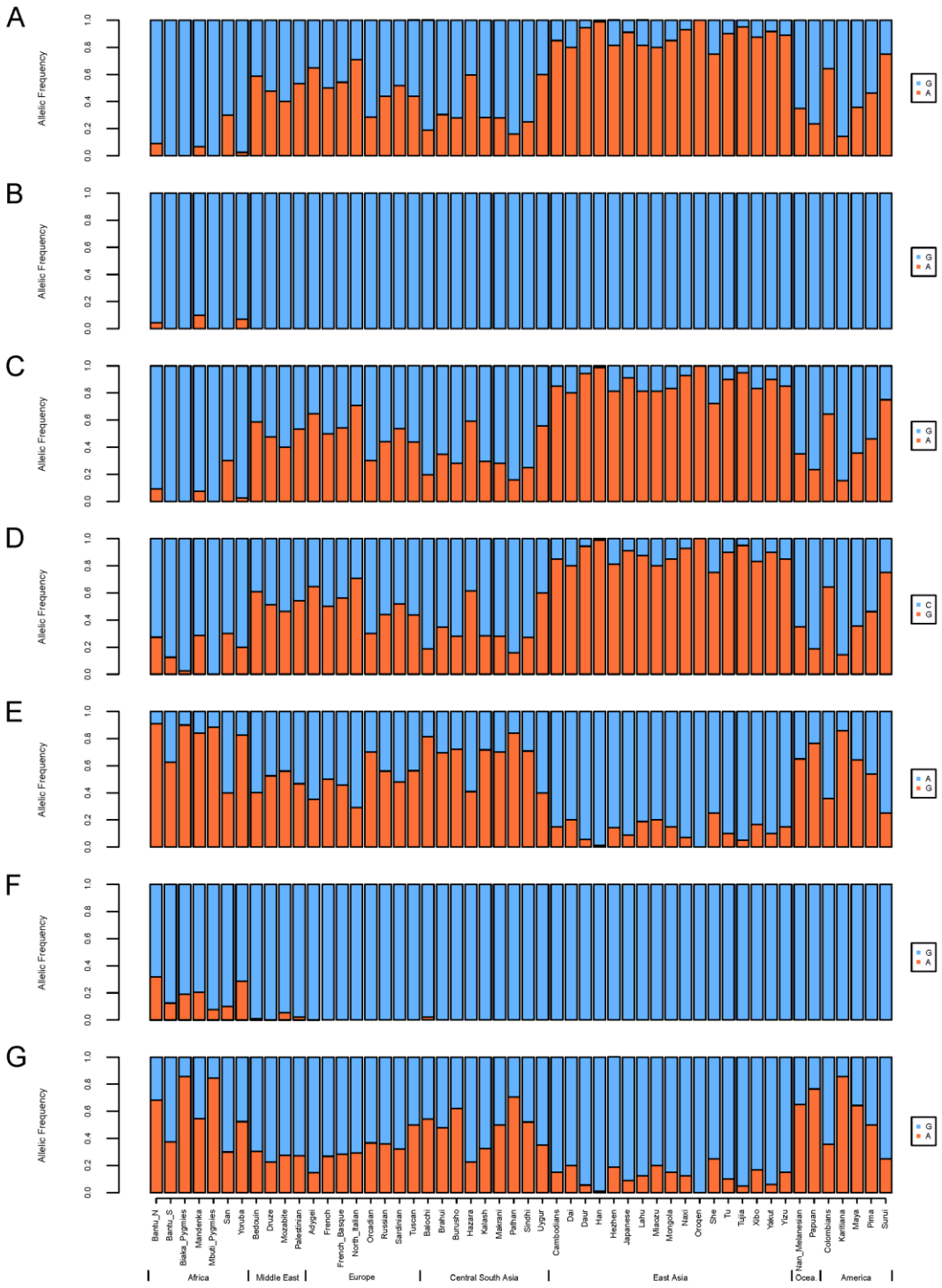


Figure S6. Allele frequency distribution of the seven *VKORC1* SNPs in the 52 HGDP-CEPH samples: rs9923231 **(A)**, rs13336384, **(B)** rs9934438 **(C)**, rs8050894 **(D)**, rs2359612 **(E)**, rs7200749 **(F)** and rs7294 **(G)**. The derived and ancestral alleles are represented in orange and blue, respectively.

Table S1. Global F_{ST} values among populations and among regions for the seven *VKORC1* SNPs.

	Global genetic differentiation among populations		Global genetic differentiation among regions	
SNP	F_{ST}[a]	*p*-value[b]	F_{ST}[c]	*p*-value[b]
rs7294	0.198	0.068	0.186	0.097
rs7200749	0.171	0.265	0.179	0.219
rs2359612	0.266	0.013*	0.270	0.016*
rs8050894	0.273	0.012*	0.277	0.015*
rs9934438	0.309	0.0065**	0.319	0.0079**
rs13336384	0.108	0.045*	0.047	0.256
rs9923231	0.309	0.0065**	0.319	0.0079**

[a] F_{ST} estimated at the global level among the 52 populations.

[b] *P*-values are derived from the genome-wide empirical distribution of F_{ST} values.

[c] F_{ST} estimated at the global level among the seven regions.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$

Table S2. Results of the XP-CLR test in a 16 kb region centered on *VKORC1* in the 52 HGDP-CEPH samples.

Region	Population	Physical position	XP-CLR score	XP-CLR *p*-value[a]
Africa	San	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.42	0.234
		31017354	0.06	0.337
		31021354	0.03	0.354
	Mbuti Pygmies	31005354	6.09	0.067
		31009354	2.68	0.157
		31013354	5.33	0.079
		31017354	6.50	0.062
		31021354	5.46	0.077
	Biaka Pygmies	31005354	6.47	0.071
		31009354	3.60	0.124
		31013354	7.33	0.061
		31017354	8.97	0.050 *
		31021354	9.04	0.049 *
	Yoruba	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.30	0.339
		31017354	0.48	0.298
		31021354	0.08	0.413
	Mandenka	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.20	0.372
		31017354	0.45	0.310
		31021354	0.06	0.427
	Bantu North	31005354	0.20	0.308
		31009354	0.17	0.317
		31013354	0.92	0.175
		31017354	0.99	0.167
		31021354	1.06	0.160
	Bantu South	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	1.31	0.115
		31017354	0.44	0.224
		31021354	0.08	0.322
Middle East	Mozabite	31005354	0.53	0.300
		31009354	0.31	0.346
		31013354	3.29	0.106
		31017354	0.43	0.319
		31021354	2.74	0.126
	Bedouin	31005354	7.18	0.054
		31009354	1.07	0.261
		31013354	8.32	0.044 *
		31017354	12.59	0.024 *
		31021354	14.44	0.018 *
	Palestinian	31005354	3.72	0.110
		31009354	0.85	0.284
		31013354	6.91	0.053
		31017354	7.53	0.048 *
		31021354	10.58	0.028 *
	Druze	31005354	0.00	0.470
		31009354	0.07	0.432
		31013354	5.08	0.089
		31017354	1.30	0.241
		31021354	5.61	0.080
Europe	Sardinian	31005354	0.79	0.301
		31009354	0.00	1.000
		31013354	4.40	0.104
		31017354	3.39	0.137
		31021354	4.88	0.093
	Tuscan	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.74	0.235
		31017354	0.08	0.352
		31021354	0.49	0.266

Table S2. (suite)

	North Italian	31005354	4.09	0.083
		31009354	0.41	0.301
		31013354	10.46	0.025 *
		31017354	11.81	0.020 *
		31021354	12.57	0.017 *
	French	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	4.19	0.117
		31017354	2.38	0.182
		31021354	4.46	0.111
	Orcadian	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.02	0.421
		31021354	0.00	0.440
	French Basque	31005354	0.52	0.306
		31009354	0.00	1.000
		31013354	4.51	0.094
		31017354	4.71	0.090
		31021354	5.36	0.080
	Russian	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	2.69	0.132
		31017354	0.26	0.343
		31021354	1.85	0.171
	Adygei	31005354	5.54	0.066
		31009354	1.08	0.227
		31013354	6.01	0.060
		31017354	10.12	0.033 *
		31021354	12.32	0.026 *
Central South Asia	Makrani	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.05	0.408
		31021354	0.00	0.444
	Balochi	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.00	1.000
		31021354	0.00	1.000
	Brahui	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.14	0.373
		31021354	0.68	0.267
	Kalash	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.03	0.423
		31021354	0.00	0.444
	Burusho	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.04	0.402
		31021354	0.00	0.431
	Pathan	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.00	1.000
		31021354	0.00	1.000
	Sindhi	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.01	0.417
		31021354	0.00	0.433
	Hazara	31005354	3.88	0.102

Table S2. (suite)

		31009354	0.58	0.282
		31013354	5.68	0.074
		31017354	8.05	0.053
		31021354	9.63	0.044 *
	Uygur	31005354	0.00	0.369
		31009354	0.00	1.000
		31013354	3.61	0.082
		31017354	4.97	0.059
		31021354	4.76	0.062
East Asia	Yakut	31005354	42.81	0.006 **
		31009354	26.29	0.019 *
		31013354	36.86	0.009 **
		31017354	35.75	0.010 **
		31021354	66.80	0.002 ***
	Mongola	31005263	21.75	0.013 *
		31009263	12.20	0.034 *
		31013263	20.17	0.015 *
		31017263	20.04	0.016 *
		31021263	29.72	0.006 **
	Tu	31005263	31.23	0.007 **
		31009263	18.63	0.017 *
		31013263	26.10	0.010 **
		31017263	25.66	0.010 *
		31021263	47.22	0.001 ***
	Xibo	31005263	17.03	0.020 *
		31009263	9.15	0.046 *
		31013263	17.78	0.019 *
		31017263	16.71	0.020 *
		31021263	25.33	0.011 *
	Oroqen	31005263	4.53	0.094
		31009263	4.52	0.095
		31013263	3.44	0.119
		31017263	3.44	0.119
		31021263	3.44	0.119
	Hezhen	31005263	13.17	0.025 *
		31009263	7.50	0.055
		31013263	14.89	0.019 *
		31017263	14.07	0.022 *
		31021263	20.26	0.006 **
	Daur	31005263	34.40	0.006 **
		31009263	22.84	0.012 *
		31013263	28.48	0.008 **
		31017263	27.57	0.008 **
		31021263	53.98	0.002 ***
	Yizu	31005263	20.97	0.014 *
		31009263	11.77	0.038 *
		31013263	19.63	0.016 *
		31017263	19.59	0.017 *
		31021263	28.91	0.006 **
	Naxi	31005263	27.12	0.004 ***
		31009263	16.35	0.015 *
		31013263	23.36	0.006 **
		31017263	22.27	0.007 **
		31021263	39.59	0.002 ***
	Tujia	31005263	36.98	0.005 ***
		31009263	26.55	0.008 **
		31013263	32.77	0.005 **
		31017263	29.66	0.006 **
		31021263	57.52	0.002 ***
	Han	31005263	89.83	0.003 ***
		31009263	102.81	0.002 ***
		31013263	118.52	0.000***
		31017263	105.94	0.002 ***
		31021263	111.55	0.001 ***
	She	31005263	5.42	0.092
		31009263	2.03	0.178

Table S2. (fin)

		31013263	11.34	0.044 *
		31017263	11.71	0.043 *
		31021263	11.85	0.043 *
	Miaozu	31005263	9.26	0.044 *
		31009263	6.08	0.074
		31013263	15.82	0.024 *
		31017263	15.16	0.025 *
		31021263	20.15	0.015 *
	Dai	31005354	9.29	0.055
		31009354	6.07	0.083
		31013354	15.79	0.028 *
		31017354	15.07	0.030 *
		31021354	20.39	0.018 *
	Lahu	31005263	16.62	0.020 *
		31009263	9.13	0.051
		31013263	15.37	0.025 *
		31017263	14.72	0.027 *
		31021263	23.70	0.008 **
	Japanese	31005263	42.96	0.005 ***
		31009263	25.63	0.015 *
		31013263	38.24	0.006 **
		31017263	36.72	0.007 **
		31021263	67.31	0.001 ***
	Cambodian	31005263	22.86	0.012 *
		31009263	12.80	0.027 *
		31013263	20.92	0.014 *
		31017263	20.65	0.014 *
		31021263	31.13	0.006 **
Oceania	Nan Melanesian	31005263	0.00	1.000
		31009263	0.00	1.000
		31013263	0.00	1.000
		31017263	0.02	0.467
		31021263	0.00	0.480
	Papuan	31005263	0.00	1.000
		31009263	0.00	1.000
		31013263	0.00	1.000
		31017263	0.00	1.000
		31021263	0.00	1.000
America	Maya	31005354	0.00	1.000
		31009354	0.00	1.000
		31013354	0.00	1.000
		31017354	0.07	0.523
		31021354	0.24	0.480
	Pima	31005263	0.00	1.000
		31009263	0.00	1.000
		31013263	0.79	0.351
		31017263	0.09	0.461
		31021263	0.55	0.382
	Colombian	31005263	0.00	1.000
		31009263	0.00	1.000
		31013263	2.04	0.172
		31017263	3.02	0.134
		31021263	2.31	0.159
	Karitiana	31005263	1.39	0.304
		31009263	0.49	0.390
		31013263	0.00	1.000
		31017263	0.00	1.000
		31021263	2.38	0.249
	Surui	31005263	2.05	0.232
		31009263	0.10	0.436
		31013263	6.12	0.107
		31017263	6.69	0.100
		31021263	6.22	0.105

[a] *P*-values are derived from the empirical distribution of XP-CLR scores along the chromosome 16.

Table S3. Results of the XP-EHH and iHS tests in the 52 HGDP-CEPH samples.

Region	Population	SNP	DAF[a]	XP-EHH score	XP-EHH *p*-value[b]	iHS score	iHS *p*-value[b]
Africa	San	rs7294	0.30	-1.47	0.926	0.56	0.513
		rs7200749	0.10	-1.55	0.937	NA	NA
		rs2359612	0.40	-1.46	0.924	1.29	0.193
		rs8050894	0.30	-1.55	0.937	1.29	0.193
		rs9934438	0.30	-1.55	0.937	1.29	0.193
		rs13336384	0.00	-1.54	0.935	NA	NA
		rs9923231	0.30	-1.55	0.937	-2.23	0.051
	Mbuti Pygmies	rs7294	0.85	1.42	0.072	-2.38	0.024 *
		rs7200749	0.08	1.27	0.097	2.15	0.038 *
		rs2359612	0.88	1.42	0.071	-1.42	0.142
		rs8050894	0.00	1.45	0.067	NA	NA
		rs9934438	0.00	1.45	0.067	NA	NA
		rs13336384	0.00	1.44	0.069	NA	NA
		rs9923231	0.00	1.45	0.067	NA	NA
	Biaka Pygmies	rs7294	0.86	0.36	0.372	-1.07	0.236
		rs7200749	0.19	-0.01	0.524	0.80	0.379
		rs2359612	0.90	0.42	0.348	-0.31	0.737
		rs8050894	0.02	0.45	0.335	NA	NA
		rs9934438	0.00	0.46	0.332	NA	NA
		rs13336384	0.00	0.45	0.336	NA	NA
		rs9923231	0.00	0.46	0.332	NA	NA
	Yoruba	rs7294	0.52	-1.29	0.893	-0.12	0.899
		rs7200749	0.29	-1.62	0.936	-0.66	0.487
		rs2359612	0.83	-1.57	0.931	0.11	0.911
		rs8050894	0.19	-1.67	0.941	-0.82	0.389
		rs9934438	0.02	-1.53	0.925	NA	NA
		rs13336384	0.07	-1.56	0.929	-0.68	0.474
		rs9923231	0.02	-1.43	0.913	NA	NA
	Mandenka	rs7294	0.55	-1.13	0.871	0.01	0.988
		rs7200749	0.20	-1.86	0.957	-1.18	0.200
		rs2359612	0.84	-1.75	0.949	0.03	0.971
		rs8050894	0.27	-1.90	0.959	-1.07	0.250
		rs9934438	0.07	-1.67	0.940	0.17	0.850
		rs13336384	0.09	-1.72	0.946	-0.45	0.620
		rs9923231	0.07	-1.58	0.931	0.23	0.804
	Bantu North	rs7294	0.68	-0.43	0.677	-0.01	0.990
		rs7200749	0.32	-1.39	0.915	0.07	0.934
		rs2359612	0.91	-0.85	0.806	0.77	0.401
		rs8050894	0.27	-1.06	0.856	0.81	0.376
		rs9934438	0.09	-0.95	0.834	-0.43	0.636
		rs13336384	0.05	-0.96	0.835	NA	NA
		rs9923231	0.09	-0.93	0.828	-0.43	0.636
	Bantu South	rs7294	0.38	-1.39	0.911	-0.66	0.473
		rs7200749	0.13	-1.54	0.932	-0.87	0.347
		rs2359612	0.63	-1.44	0.918	0.06	0.949
		rs8050894	0.13	-0.95	0.830	-0.97	0.297
		rs9934438	0.00	-0.84	0.804	NA	NA
		rs13336384	0.00	-0.85	0.806	NA	NA
		rs9923231	0.00	-0.84	0.804	NA	NA
Middle East	Mozabite	rs7294	0.28	1.24	0.111	1.63	0.088
		rs7200749	0.05	1.65	0.057	-0.32	0.729
		rs2359612	0.55	1.60	0.064	2.38	0.019 *
		rs8050894	0.47	1.68	0.055	-1.48	0.118
		rs9934438	0.41	1.53	0.070	-1.84	0.056
		rs13336384	0.00	1.57	0.067	NA	NA
		rs9923231	0.41	1.42	0.084	-1.84	0.056
	Bedouin	rs7294	0.30	1.44	0.086	1.71	0.074
		rs7200749	0.01	1.78	0.048 *	NA	NA
		rs2359612	0.40	1.79	0.047 *	2.73	0.010 *
		rs8050894	0.61	1.89	0.039 *	-1.59	0.094
		rs9934438	0.59	1.74	0.052	-1.76	0.067
		rs13336384	0.00	1.77	0.049 *	NA	NA
		rs9923231	0.59	1.63	0.063	-1.76	0.067
	Palestinian	rs7294	0.27	1.35	0.099	1.56	0.099
		rs7200749	0.02	1.70	0.053	2.32	0.021 *
		rs2359612	0.47	1.67	0.058	2.77	0.009 **
		rs8050894	0.54	1.77	0.047 *	-1.67	0.079
		rs9934438	0.53	1.62	0.063	-1.81	0.058
		rs13336384	0.00	1.65	0.060	NA	NA
		rs9923231	0.53	1.51	0.077	-1.81	0.058
	Druze	rs7294	0.23	0.86	0.188	1.13	0.236
		rs7200749	0.00	1.27	0.110	NA	NA
		rs2359612	0.52	1.18	0.124	1.81	0.062
		rs8050894	0.51	1.28	0.110	-0.54	0.567
		rs9934438	0.48	1.14	0.132	-0.90	0.342
		rs13336384	0.00	1.17	0.127	NA	NA
		rs9923231	0.48	1.03	0.152	-0.90	0.342
Europe	Sardinian	rs7294	0.32	1.04	0.154	0.52	0.571
		rs7200749	0.00	1.40	0.088	NA	NA
		rs2359612	0.48	1.39	0.090	1.67	0.084
		rs8050894	0.52	1.49	0.078	-0.85	0.360
		rs9934438	0.52	1.35	0.097	-0.85	0.360
		rs13336384	0.00	1.38	0.092	NA	NA
		rs9923231	0.52	1.25	0.113	-0.85	0.360
	Tuscan	rs7294	0.50	0.59	0.264	1.01	0.273
		rs7200749	0.00	0.98	0.165	NA	NA
		rs2359612	0.56	1.01	0.159	1.51	0.110
		rs8050894	0.44	1.08	0.145	-0.77	0.397
		rs9934438	0.44	0.91	0.184	-0.77	0.397
		rs13336384	0.00	0.94	0.174	NA	NA
		rs9923231	0.44	0.79	0.210	-0.77	0.397
	North Italian	rs7294	0.29	2.15	0.023 *	1.91	0.050

Table S3. (suite)

		rs7200749	0.00	2.45	0.014 *	NA	NA
		rs2359612	0.29	2.50	0.012 *	1.91	0.050
		rs8050894	0.71	2.60	0.010 *	-1.19	0.197
		rs9934438	0.71	2.45	0.013 *	-1.19	0.197
		rs13336384	0.00	2.48	0.013 *	NA	NA
		rs9923231	0.71	2.35	0.016 *	-1.19	0.197
	French	rs7294	0.27	1.29	0.107	0.68	0.470
		rs7200749	0.00	1.62	0.064	NA	NA
		rs2359612	0.50	1.57	0.069	1.87	0.055
		rs8050894	0.50	1.67	0.059	-1.00	0.289
		rs9934438	0.50	1.53	0.075	-1.00	0.289
		rs13336384	0.00	1.56	0.071	NA	NA
		rs9923231	0.50	1.43	0.088	-1.00	0.289
	Orcadian	rs7294	0.37	0.24	0.379	-0.14	0.871
		rs7200749	0.00	0.68	0.239	NA	NA
		rs2359612	0.70	0.49	0.294	0.97	0.280
		rs8050894	0.30	0.54	0.278	-0.40	0.647
		rs9934438	0.30	0.37	0.335	-0.40	0.647
		rs13336384	0.00	0.41	0.322	NA	NA
		rs9923231	0.30	0.25	0.376	-0.40	0.647
	French Basque	rs7294	0.29	1.25	0.112	0.48	0.593
		rs7200749	0.00	1.56	0.071	NA	NA
		rs2359612	0.46	1.56	0.071	1.72	0.068
		rs8050894	0.56	1.65	0.062	-0.61	0.498
		rs9934438	0.54	1.52	0.075	-0.88	0.331
		rs13336384	0.00	1.55	0.072	NA	NA
		rs9923231	0.54	1.42	0.086	-0.88	0.331
	Russian	rs7294	0.36	1.13	0.134	0.37	0.680
		rs7200749	0.00	1.47	0.081	NA	NA
		rs2359612	0.56	1.46	0.083	1.55	0.097
		rs8050894	0.44	1.55	0.071	-0.73	0.420
		rs9934438	0.44	1.41	0.090	-0.73	0.420
		rs13336384	0.00	1.44	0.087	NA	NA
		rs9923231	0.44	1.30	0.105	-0.73	0.420
	Adygei	rs7294	0.15	2.25	0.020 *	0.959	0.290
		rs7200749	0.00	2.55	0.010 *	NA	NA
		rs2359612	0.35	2.53	0.011 *	2.231	0.025 *
		rs8050894	0.65	2.63	0.009 **	-1.331	0.152
		rs9934438	0.65	2.49	0.012 *	-1.331	0.152
		rs13336384	0.00	2.51	0.011 *	NA	NA
		rs9923231	0.65	2.38	0.015 *	-1.331	0.152
Central South Asia	Makrani	rs7294	0.50	0.35	0.348	-0.71	0.441
		rs7200749	0.00	0.84	0.192	NA	NA
		rs2359612	0.70	0.71	0.228	0.36	0.695
		rs8050894	0.28	0.77	0.211	-0.01	0.995
		rs9934438	0.28	0.59	0.265	-0.01	0.995
		rs13336384	0.00	0.62	0.254	NA	NA
		rs9923231	0.28	0.46	0.305	-0.01	0.995
	Balochi	rs7294	0.54	0.94	0.168	-0.60	0.484
		rs7200749	0.02	1.31	0.106	NA	NA
		rs2359612	0.81	1.29	0.109	0.56	0.516
		rs8050894	0.19	1.40	0.095	-0.04	0.967
		rs9934438	0.19	1.25	0.114	-0.04	0.967
		rs13336384	0.00	1.27	0.111	NA	NA
		rs9923231	0.19	1.14	0.131	-0.04	0.967
	Brahui	rs7294	0.48	0.42	0.322	-0.20	0.820
		rs7200749	0.00	0.94	0.177	NA	NA
		rs2359612	0.66	0.81	0.206	0.96	0.285
		rs8050894	0.34	0.88	0.190	-0.11	0.900
		rs9934438	0.34	0.70	0.234	-0.11	0.900
		rs13336384	0.00	0.74	0.225	NA	NA
		rs9923231	0.34	0.57	0.271	-0.11	0.900
	Kalash	rs7294	0.33	0.38	0.350	0.21	0.809
		rs7200749	0.00	0.82	0.212	NA	NA
		rs2359612	0.72	0.73	0.237	0.38	0.670
		rs8050894	0.28	0.83	0.208	0.18	0.842
		rs9934438	0.28	0.68	0.252	0.18	0.842
		rs13336384	0.00	0.71	0.244	NA	NA
		rs9923231	0.28	0.58	0.284	0.18	0.842
	Burusho	rs7294	0.62	2.10	0.027 *	-1.25	0.181
		rs7200749	0.00	2.45	0.013 *	NA	NA
		rs2359612	0.72	2.48	0.012 *	-0.34	0.706
		rs8050894	0.28	2.58	0.010 *	1.06	0.250
		rs9934438	0.28	2.44	0.013 *	1.06	0.250
		rs13336384	0.00	2.46	0.013 *	NA	NA
		rs9923231	0.28	2.33	0.017 *	1.06	0.250
	Pathan	rs7294	0.70	1.77	0.052	-1.06	0.259
		rs7200749	0.00	2.13	0.026 *	NA	NA
		rs2359612	0.84	2.18	0.023 *	0.33	0.728
		rs8050894	0.16	2.29	0.018 *	0.17	0.856
		rs9934438	0.16	2.14	0.026 *	0.17	0.856
		rs13336384	0.00	2.16	0.024 *	NA	NA
		rs9923231	0.16	2.03	0.032 *	0.17	0.856
	Sindhi	rs7294	0.52	1.00	0.165	-1.08	0.239
		rs7200749	0.00	1.37	0.099	NA	NA
		rs2359612	0.71	1.37	0.099	0.12	0.894
		rs8050894	0.27	1.47	0.083	0.45	0.624
		rs9934438	0.25	1.32	0.106	0.14	0.874
		rs13336384	0.00	1.35	0.102	NA	NA
		rs9923231	0.25	1.21	0.124	0.14	0.874
	Hazara	rs7294	0.23	1.74	0.051	1.40	0.141
		rs7200749	0.00	2.08	0.027 *	NA	NA
		rs2359612	0.41	2.05	0.028 *	2.37	0.018 *
		rs8050894	0.61	2.15	0.023 *	-1.28	0.178

Table S3. (suite)

		rs9934438	0.59	2.01	0.030 *	-1.51	0.115
		rs13336384	0.00	2.04	0.029 *	NA	NA
		rs9923231	0.59	1.91	0.038 *	-1.51	0.115
	Uygur	rs7294	0.35	2.23	0.022 *	0.39	0.673
		rs7200749	0.00	2.53	0.012 *	NA	NA
		rs2359612	0.40	2.56	0.012 *	1.42	0.134
		rs8050894	0.60	2.67	0.010 *	1.42	0.134
		rs9934438	0.60	2.52	0.013 *	1.42	0.134
		rs13336384	0.00	2.55	0.012 *	NA	NA
		rs9923231	0.60	2.48	0.013 *	-0.90	0.327
East Asia	Yakut	rs7294	0.06	3.17	0.007 **	1.44	0.124
		rs7200749	0.00	3.41	0.005 **	NA	NA
		rs2359612	0.10	3.45	0.005 **	3.15	0.007 **
		rs8050894	0.90	3.53	0.005 **	-2.39	0.018 *
		rs9934438	0.90	3.41	0.005 **	-2.39	0.018 *
		rs13336384	0.00	3.43	0.005 **	NA	NA
		rs9923231	0.90	3.33	0.006 **	-2.39	0.018 *
	Mongola	rs7294	0.15	2.76	0.014 *	1.30	0.172
		rs7200749	0.00	3.00	0.011 *	NA	NA
		rs2359612	0.15	3.05	0.010 *	1.30	0.172
		rs8050894	0.85	3.13	0.009 **	-1.30	0.172
		rs9934438	0.85	3.02	0.011 *	-1.30	0.172
		rs13336384	0.00	3.03	0.010 *	NA	NA
		rs9923231	0.85	2.93	0.012 *	-1.30	0.172
	Tu	rs7294	0.10	3.29	0.006 **	0.47	0.617
		rs7200749	0.00	3.54	0.003 **	NA	NA
		rs2359612	0.10	3.61	0.003 **	0.47	0.617
		rs8050894	0.90	3.70	0.003 **	0.04	0.971
		rs9934438	0.90	3.57	0.003 **	0.04	0.971
		rs13336384	0.00	3.59	0.003 **	NA	NA
		rs9923231	0.90	3.48	0.004 **	0.04	0.971
	Xibo	rs7294	0.17	2.73	0.008 **	1.64	0.089
		rs7200749	0.00	2.99	0.005 **	NA	NA
		rs2359612	0.17	3.08	0.004 **	1.64	0.089
		rs8050894	0.83	3.17	0.003 **	-0.94	0.312
		rs9934438	0.83	3.05	0.004 **	-0.94	0.312
		rs13336384	0.00	3.06	0.004 **	NA	NA
		rs9923231	0.83	2.96	0.006 **	-0.94	0.312
	Oroqen	rs7294	0.00	3.15	0.006 **	NA	NA
		rs7200749	0.00	3.39	0.004 **	NA	NA
		rs2359612	0.00	3.47	0.003 **	NA	NA
		rs8050894	1.00	3.56	0.003 **	NA	NA
		rs9934438	1.00	3.44	0.004 **	NA	NA
		rs13336384	0.00	3.45	0.004 **	NA	NA
		rs9923231	1.00	3.35	0.004 **	NA	NA
	Hezhen	rs7294	0.19	2.51	0.013 *	2.00	0.045 *
		rs7200749	0.00	2.76	0.008 **	NA	NA
		rs2359612	0.19	2.82	0.007 **	2.00	0.045 *
		rs8050894	0.81	2.91	0.006 **	-1.40	0.137
		rs9934438	0.81	2.79	0.008 **	-1.40	0.137
		rs13336384	0.00	2.80	0.007 **	NA	NA
		rs9923231	0.81	2.70	0.009 **	-1.40	0.137
	Daur	rs7294	0.06	2.99	0.008 **	NA	NA
		rs7200749	0.00	3.24	0.006 **	NA	NA
		rs2359612	0.06	3.30	0.005 **	NA	NA
		rs8050894	0.94	3.38	0.005 **	NA	NA
		rs9934438	0.94	3.26	0.006 **	NA	NA
		rs13336384	0.00	3.28	0.006 **	NA	NA
		rs9923231	0.94	3.18	0.007 **	NA	NA
	Yizu	rs7294	0.15	2.83	0.009 **	0.35	0.698
		rs7200749	0.00	3.10	0.005 **	NA	NA
		rs2359612	0.15	3.18	0.004 **	0.35	0.698
		rs8050894	0.85	3.27	0.003 **	-0.35	0.698
		rs9934438	0.85	3.14	0.004 **	-0.35	0.698
		rs13336384	0.00	3.16	0.004 **	NA	NA
		rs9923231	0.85	3.06	0.005 **	-0.35	0.698
	Naxi	rs7294	0.13	2.91	0.004 **	0.74	0.396
		rs7200749	0.00	3.17	0.002 **	NA	NA
		rs2359612	0.13	3.27	0.001 **	0.74	0.396
		rs8050894	0.88	3.35	0.001 **	-0.21	0.811
		rs9934438	0.88	3.23	0.002 **	-0.21	0.811
		rs13336384	0.00	3.25	0.001 **	NA	NA
		rs9923231	0.88	3.14	0.002 **	-0.21	0.811
	Tujia	rs7294	0.05	3.40	0.006 **	NA	NA
		rs7200749	0.00	3.64	0.005 **	NA	NA
		rs2359612	0.05	3.69	0.004 **	NA	NA
		rs8050894	0.95	3.78	0.004 **	NA	NA
		rs9934438	0.95	3.66	0.005 **	NA	NA
		rs13336384	0.00	3.67	0.004 **	NA	NA
		rs9923231	0.95	3.57	0.005 **	NA	NA
	Han	rs7294	0.01	3.23	0.005 **	NA	NA
		rs7200749	0.00	3.47	0.003 **	NA	NA
		rs2359612	0.01	3.51	0.003 **	NA	NA
		rs8050894	0.99	3.59	0.002 **	NA	NA
		rs9934438	0.99	3.47	0.003 **	NA	NA
		rs13336384	0.00	3.49	0.003 **	NA	NA
		rs9923231	0.99	3.39	0.004 **	NA	NA
	She	rs7294	0.25	2.08	0.033 *	1.17	0.203
		rs7200749	0.00	2.33	0.020 *	NA	NA
		rs2359612	0.25	2.40	0.017 *	1.17	0.203
		rs8050894	0.75	2.49	0.014 *	-0.59	0.503
		rs9934438	0.75	2.37	0.019 *	-0.59	0.503
		rs13336384	0.00	2.38	0.018 *	NA	NA
		rs9923231	0.75	2.30	0.022 *	-0.59	0.503
	Miaozu	rs7294	0.20	2.26	0.020 *	0.68	0.205

Table S3. (fin)

		rs7200749	0.00	2.53	0.014 *	NA	NA
		rs2359612	0.20	2.61	0.012 *	0.68	0.205
		rs8050894	0.80	2.70	0.011 *	-0.27	0.604
		rs9934438	0.80	2.57	0.013 *	-0.27	0.604
		rs13336384	0.00	2.59	0.012 *	NA	NA
		rs9923231	0.80	2.48	0.015 *	-0.27	0.604
	Dai	rs7294	0.20	1.84	0.049 *	0.90	0.319
		rs7200749	0.00	2.11	0.030 *	NA	NA
		rs2359612	0.20	2.16	0.028 *	0.90	0.319
		rs8050894	0.80	2.26	0.022 *	-0.43	0.628
		rs9934438	0.80	2.13	0.029 *	-0.43	0.628
		rs13336384	0.00	2.15	0.028 *	NA	NA
		rs9923231	0.80	2.04	0.035 *	-0.43	0.628
	Lahu	rs7294	0.13	1.97	0.036 *	1.04	0.246
		rs7200749	0.00	2.24	0.023 *	NA	NA
		rs2359612	0.19	2.27	0.022 *	2.54	0.019 *
		rs8050894	0.88	2.37	0.017 *	-0.55	0.528
		rs9934438	0.81	2.24	0.023 *	-2.13	0.034 *
		rs13336384	0.00	2.26	0.022 *	NA	NA
		rs9923231	0.81	2.15	0.027 *	-2.13	0.034 *
	Japanese	rs7294	0.09	2.94	0.008 **	1.32	0.166
		rs7200749	0.00	3.18	0.006 **	NA	NA
		rs2359612	0.09	3.22	0.005 **	1.32	0.166
		rs8050894	0.91	3.30	0.005 **	-0.93	0.319
		rs9934438	0.91	3.19	0.006 **	-0.93	0.319
		rs13336384	0.00	3.20	0.006 **	NA	NA
		rs9923231	0.91	3.10	0.007 **	-0.93	0.319
	Cambodian	rs7294	0.15	2.53	0.012 *	-0.09	0.921
		rs7200749	0.00	2.81	0.007 **	NA	NA
		rs2359612	0.15	2.90	0.006 **	-0.09	0.921
		rs8050894	0.85	3.00	0.005 **	0.09	0.921
		rs9934438	0.85	2.86	0.006 **	0.09	0.921
		rs13336384	0.00	2.88	0.006 **	NA	NA
		rs9923231	0.85	2.77	0.008 **	0.09	0.921
Oceania	Nan Melanesian	rs7294	0.65	-0.19	0.535	0.17	0.851
		rs7200749	0.00	0.11	0.410	NA	NA
		rs2359612	0.65	0.15	0.395	0.17	0.851
		rs8050894	0.35	0.21	0.374	0.17	0.851
		rs9934438	0.35	0.07	0.424	0.17	0.851
		rs13336384	0.00	0.10	0.414	NA	NA
		rs9923231	0.35	-0.02	0.462	0.02	0.982
	Papuan	rs7294	0.76	0.12	0.400	-0.08	0.932
		rs7200749	0.00	0.34	0.321	NA	NA
		rs2359612	0.76	0.40	0.301	-0.08	0.932
		rs8050894	0.24	0.41	0.297	0.47	0.609
		rs9934438	0.24	0.30	0.333	0.47	0.609
		rs13336384	0.00	0.32	0.327	NA	NA
		rs9923231	0.24	0.22	0.361	0.47	0.609
America	Maya	rs7294	0.64	1.42	0.090	-0.06	0.948
		rs7200749	0.00	1.65	0.062	NA	NA
		rs2359612	0.64	1.70	0.058	-0.06	0.948
		rs8050894	0.36	1.77	0.052	0.83	0.381
		rs9934438	0.36	1.66	0.062	0.83	0.381
		rs13336384	0.00	1.68	0.060	NA	NA
		rs9923231	0.36	1.58	0.071	0.83	0.381
	Pima	rs7294	0.50	0.76	0.209	NA	NA
		rs7200749	0.00	0.96	0.172	NA	NA
		rs2359612	0.50	0.99	0.165	NA	NA
		rs8050894	0.50	1.06	0.154	NA	NA
		rs9934438	0.50	0.96	0.170	NA	NA
		rs13336384	0.00	0.98	0.168	NA	NA
		rs9923231	0.50	0.89	0.183	NA	NA
	Colombian	rs7294	0.36	0.77	0.211	NA	NA
		rs7200749	0.00	1.04	0.146	NA	NA
		rs2359612	0.36	1.25	0.108	NA	NA
		rs8050894	0.64	0.93	0.171	NA	NA
		rs9934438	0.64	0.79	0.206	NA	NA
		rs13336384	0.00	0.81	0.200	NA	NA
		rs9923231	0.64	0.69	0.230	NA	NA
	Karitiana	rs7294	0.86	NA	NA	NA	NA
		rs7200749	0.00	NA	NA	NA	NA
		rs2359612	0.86	NA	NA	NA	NA
		rs8050894	0.14	NA	NA	NA	NA
		rs9934438	0.14	NA	NA	NA	NA
		rs13336384	0.00	NA	NA	NA	NA
		rs9923231	0.14	NA	NA	NA	NA
	Surui	rs7294	0.25	0.33	0.338	NA	NA
		rs7200749	0.00	0.54	0.268	NA	NA
		rs2359612	0.25	0.61	0.247	NA	NA
		rs8050894	0.75	0.58	0.256	NA	NA
		rs9934438	0.75	0.47	0.290	NA	NA
		rs13336384	0.00	0.49	0.285	NA	NA
		rs9923231	0.75	0.39	0.317	NA	NA

[a] Derived allele frequency.

[b] *P*-values are derived from the empirical distribution of the iHS and XP-EHH scores along the chromosome 16.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$.

NA: Not Applicable. (for iHS: when a gap > 200 kb between successive SNPs is found in the region in the region delimited by the SNPs where the EHH value drops below 0.05 around the core SNP; for XP-EHH: when no SNP with an EHH of between 0.03 and 0.05 is found up to 1 Mb from the core SNP).

Table S4. Results of the XP-CLR test in the ~ 500 kb genomic region of the LD block encompassing *VKORC1* in East Asia.

Physical position	XP-CLR score	XP-CLR *p*-value[a]	Genes present in the genomic window [b]
30729354	41.16	0.01425*	*MGC2474, ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK*
30733354	40.28	0.01491*	*MGC2474, ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1*
30737354	22.98	0.03413*	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30741354	3.06	0.20347	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30745354	5.15	0.14266	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30749354	8.14	0.10117	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30753354	5.49	0.13633	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30757354	2.98	0.20647	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30761354	3.49	0.18912	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30765354	5.40	0.13751	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30769354	7.83	0.10427	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30773354	8.00	0.10267	*ZNF688, FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30777354	6.33	0.12386	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30781354	2.82	0.21341	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30785354	1.53	0.28955	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30789354	3.09	0.20230	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30793354	4.56	0.15720	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30797354	3.27	0.19616	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30801354	1.37	0.30342	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30805354	14.12	0.06048	*FLJ32130, PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30809354	21.12	0.03746*	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30813354	15.76	0.05270	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30817354	39.01	0.01613*	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30821354	54.42	0.00759**	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30825354	49.08	0.00933**	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30829354	43.10	0.01266*	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30833354	10.36	0.08008	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30837354	0.70	0.37037	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30841354	2.56	0.22527	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30845354	5.83	0.13085	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30849354	3.90	0.17543	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30853354	1.05	0.33047	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30857354	0.61	0.38200	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30861354	4.61	0.15598	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30865354	6.26	0.12457	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30869354	7.42	0.10919	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30873354	5.94	0.12916	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30877354	1.55	0.28729	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30881354	1.79	0.26920	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30885354	1.79	0.26929	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30889354	5.15	0.14252	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30893354	6.65	0.11913	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30897354	9.71	0.08537	*PRR14, SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30901354	14.13	0.06043	*SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30905354	18.17	0.04580*	*SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30909354	7.00	0.11481	*SRCAP, RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30913354	2.98	0.20661	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30917354	24.15	0.03221*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30921354	26.25	0.02930*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30925354	62.62	0.00511**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30929354	67.20	0.00455***	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30933354	59.86	0.00591**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30937354	67.32	0.00450***	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30941354	69.57	0.00422***	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30945354	40.12	0.01519*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30949354	33.45	0.02035*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30953354	13.56	0.06273	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30957354	10.79	0.07754	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30961354	15.18	0.05537	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30965354	34.84	0.01908*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30969354	44.89	0.01116*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30973354	48.57	0.00956**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30977354	58.55	0.00619**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30981354	60.85	0.00549**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30985354	56.94	0.00661**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30989354	50.66	0.00858**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30993354	44.56	0.01153*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
30997354	43.68	0.01214*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31001354	35.85	0.01824*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31005354	24.08	0.03240*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31009354	16.53	0.04993*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31013354	30.49	0.02311*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31017354	26.82	0.02822*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31021354	43.44	0.01228*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31025354	51.88	0.00811**	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31029354	3.23	0.19761	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8*
31033354	6.04	0.12780	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS*
31037354	6.12	0.12677	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS*
31041354	11.60	0.07267	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS*
31045354	7.80	0.10450	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS*
31049354	21.29	0.03704*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS, PYCARD*
31053354	5.27	0.13980	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS, PYCARD*
31057354	44.86	0.01120*	*RNF40, BCL7C, NCRNA00095, FBXL19, SETD1A, STX1B2, STX4A, VKORC1, BCKDK, MYST1, PRSS8, FUS, PYCARD*

[a] *P*-values are derived from the empirical distribution of the XP-CLR scores along the chromosome 16.

[b] Genes are named according to the NCBI 36.3 assembly.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$.

Table S5. Results of the XP-EHH, iHS tests, inter-regional F_{ST} and global F_{ST} for all SNPs located in the linkage disequilibrium block encompassing *VKORC1* in East Asia.

SNP	Physical position	GeneSymbol[a]	Location	XP-EHH *p*-value[b]	iHS *p*-value[b]	interregional F_{ST} *p*-value[c,d]	global F_{ST} *p*-value[d,e]
rs7197475	30550368	*PRR14*	flanking_5UTR	0.025*	0.235	0.097	0.116
rs3747481	30573868	*PRR14*	coding	0.027*	0.068	0.231	0.313
rs893924	30582547	*FBRS*	flanking_5UTR	0.034*	NA	0.698	0.519
rs8058578	30633749	*SRCAP*	intron	0.035*	0.065	0.230	0.303
rs11150595	30659704	*SRCAP*	flanking_3UTR	0.033*	0.065	0.122	0.014*
rs11642466	30689443	*RNF40*	intron	0.060	NA	0.502	0.827
rs8058961	30716564	*RNF40*	flanking_3UTR	0.037*	0.051	0.229	0.339
rs4889490	30730548	*RNF40*	flanking_3UTR	0.071	0.140	0.012*	0.0036***
rs8046001	30740822	*RNF40*	flanking_3UTR	0.035*	0.053	0.075	0.093
rs11150596	30757743	*BCL7C*	flanking_3UTR	0.033*	0.053	0.100	0.111
rs4889630	30785045	*BCL7C*	flanking_3UTR	0.044*	0.050	0.429	0.250
rs4889651	30803046	*BCL7C*	flanking_3UTR	0.050	0.048*	0.271	0.170
rs9933843	30811180	*BCL7C*	intron	0.061	0.190	0.048*	0.007**
rs12924903	30836471	*NCRNA00095*	flanking_3UTR	0.031*	0.027*	0.124	0.204
rs7200879	30855073	*FBXL19*	intron	0.043*	0.036*	0.259	0.088
rs2305884	30878242	*SETD1A*	intron	0.019*	0.064	0.071	0.115
rs897986	30888403	*SETD1A*	intron	0.018*	0.042*	0.077	0.135
rs12445568	30912313	*STX1B*	intron	0.024*	0.136	0.078	0.148
rs10871454	30955580	*STX4*	intron	0.049*	0.177	0.0034***	0.0093**
rs4889533	30986508	*ZNF668*	intron	0.046*	NA	0.532	0.838
rs7294	31009822	*VKORC1*	3UTR	0.011*	0.041*	0.063	0.097
rs7200749	31010090	*VKORC1*	coding	0.0053**	NA	0.576	0.219
rs2359612	31011297	*VKORC1*	intron	0.0077**	0.047*	0.0049***	0.016*
rs8050894	31012010	*VKORC1*	intron	0.0079**	0.2	0.0052**	0.015*
rs9934438	31012379	*VKORC1*	intron	0.0088**	0.174	0.0030***	0.0079**
rs13336384	31012672	*VKORC1*	intron	0.0087**	NA	0.252	0.255
rs9923231	31015190	*VKORC1*	flanking_5UTR	0.0097**	0.174	0.0030***	0.0079**
rs889555	31030072	*BCKDK*	intron	0.0056**	0.041*	0.102	0.105
rs749767	31031908	*BCKDK*	flanking_3UTR	0.0054**	0.200	0.0062**	0.021*
rs1978487	31037443	*MYST1*	intron	0.0052**	0.055	0.011*	0.048*
rs11865499	31039751	*MYST1*	intron	0.0071**	0.072	0.087	0.087
rs889548	31045213	*MYST1*	intron	0.0088**	0.204	0.0039***	0.013*
rs2855475	31055049	*PRSS8*	flanking_5UTR	0.0093**	0.056	0.0069**	0.035*

[a] Genes are named according to the NCBI 36.3 assembly.

[b] *P*-values are derived from the empirical distribution of the XP-EHH and iHS scores along the chromosome 16.

[c] F_{ST} estimated at the interregional level for East Asia, *i.e.* considering all East Asian samples versus all the remaining ones.

[d] *P*-values are derived from the genome-wide empirical distribution of F_{ST} values.

[e] F_{ST} estimated at the global level, *i.e.* among the seven regions.

* $p < 0.05$; ** $p < 0.01$; *** $p < 0.005$.

NA: Not Applicable (for iHS: when a gap > 200 kb between successive SNPs is found in the region in the region delimited by the SNPs where the EHH value drops below 0.05 around the core SNP).

Table S6. Description of the 52 HGDP-CEPH samples grouped into seven main geographic regions.

Geographic origin	Population	Sample size[a]
Africa		101
Namibia	San	5
Democratic Republic of Congo	Mbuti Pygmies	13
Central African Republic	Biaka Pygmies	21
Nigeria	Yoruba	21
Senegal	Mandenka	22
Kenya	Bantu North	11
South Africa	Bantu South	8
Middle East		163
Algeria	Mozabite	29
Israel	Bedouin	46
Israel	Palestinian	46
Israel	Druze	42
Europe		157
Italy	Sardinian	28
Italy	Tuscan	8
Italy	North Italian	12
France	French	28
Orkney Islands	Orcadian	15
France	French Basque	24
Russia	Russian	25
Russia Caucasus	Adygei	17
Central South Asia		200
Pakistan	Makrani	25
Pakistan	Balochi	24
Pakistan	Brahui	25
Pakistan	Kalash	23
Pakistan	Burusho	25
Pakistan	Pathan	22
Pakistan	Sindhi	24
Pakistan	Hazara	22
China	Uygur	10
East Asia		228
Siberia	Yakut	25
China	Mongola	10
China	Tu	10
China	Xibo	9
China	Oroqen	9
China	Hezhen	8
China	Daur	9
China	Yizu	10
China	Naxi	8
China	Tujia	10
China	Han	44
China	She	10
China	Miaozu	10
China	Dai	10
China	Lahu	8
Japan	Japanese	28
Cambodia	Cambodian	10
Oceania		27
Bougainville	Nan Melanesian	10
NewGuinea	Papuan	17
America		64
Mexico	Maya	21
Mexico	Pima	14
Colombia	Colombian	7
Brazil	Karitiana	14
Brazil	Surui	8

[a] Number of individuals.

Annexe 2

Tables et Figures supplémentaires de l'article 2

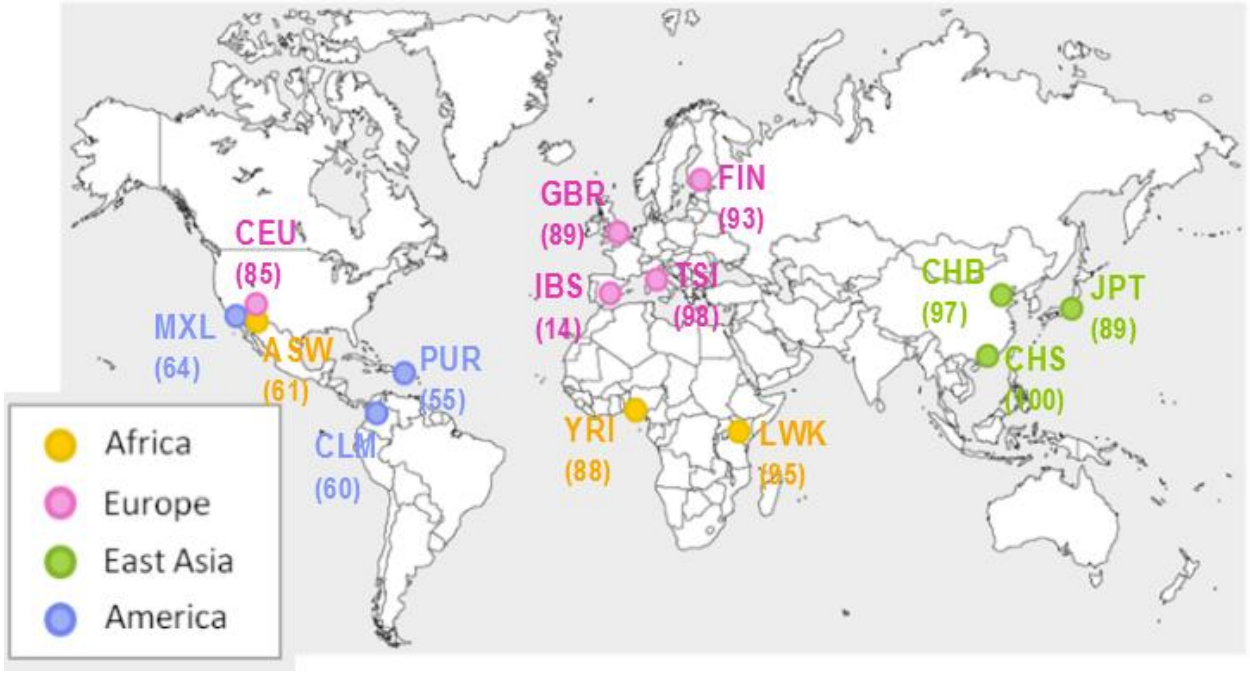


Figure S1. Geographical distribution of the 14 population samples included in the 1000 Genomes project. Population samples from Africa, Europe, East Asia and the Americas are represented by yellow, pink, green and blue dots, respectively. The assignment of populations to one of the four continental regions was based on the origin of the population, ignoring the past 1,000 years of known human migration (e.g., people of European descent in the United States were assigned to Europe). The sample size (number of individuals) is indicated in brackets. YRI: Yoruba from Ibadan, Nigeria; LWK: Luhya fromWebuye, Kenya; ASW: people of African ancestry from the southwestern United States; IBS: Iberian populations from Spain; TSI: Toscani from Italy; CEU: Utah residents with Northern and Westhern European ancestry; GBR: British from England and Scotland; FIN: Finnish from Finland; JPT: Japanese from Tokyo, Japan; CHB: Han Chinese from Beijing; CHS: Han Chinese from South China; CLM: Colombians from Medellín, Colombia; MXL: people of Mexican ancestry from Los Angeles, California; PUR: Puerto Ricans from Puerto Rico.

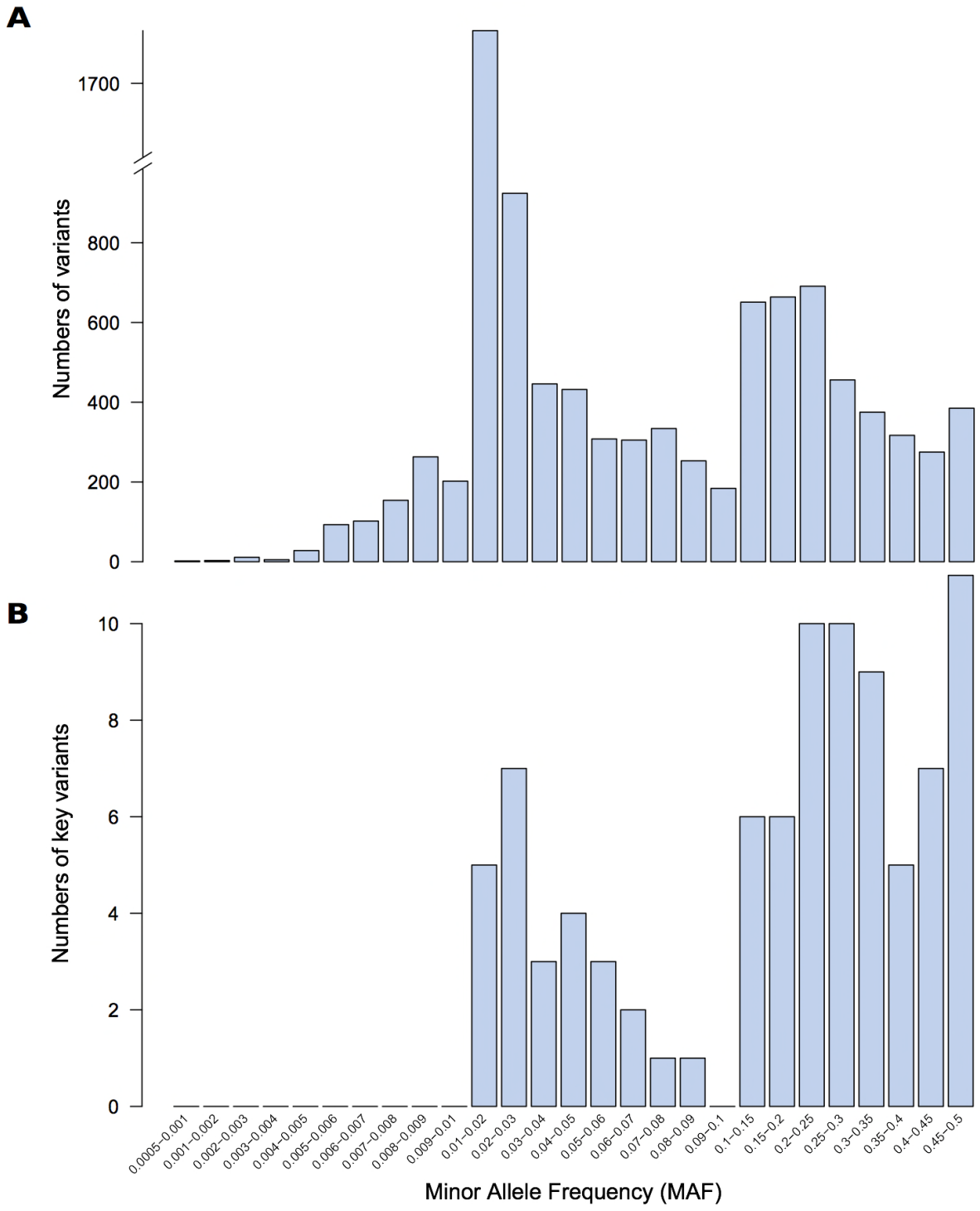


Figure S2. Distribution of variants in the 45 VIP genes according to the minor allele frequency (MAF). (A) Whole set of 9695 SNVs. (B) Subset of 90 key variants.

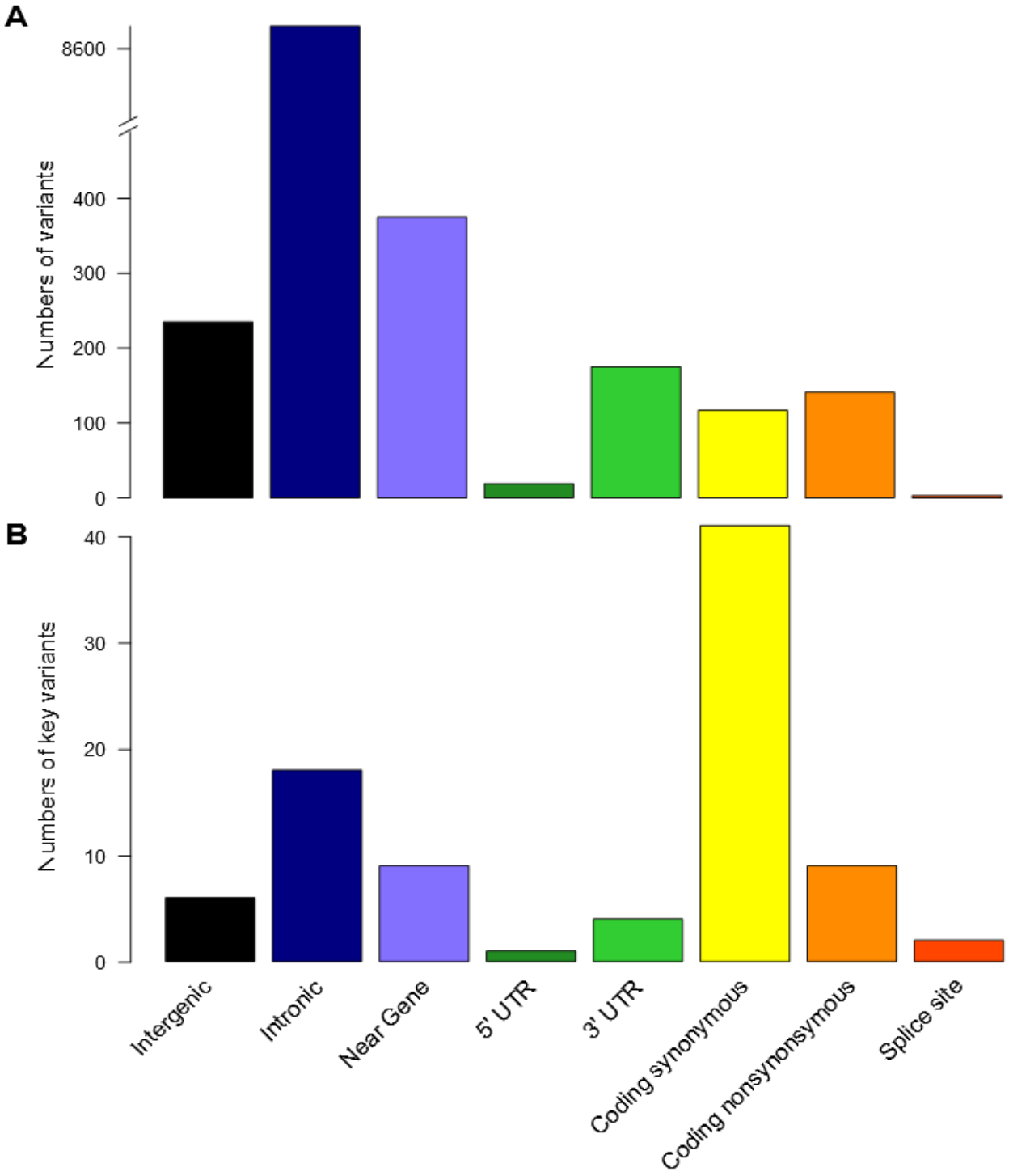


Figure S3. Distribution of variants in the 45 VIP genes according to the functional annotation class. (A) Whole set of 9695 SNVs. (B) Subset of 90 key variants.

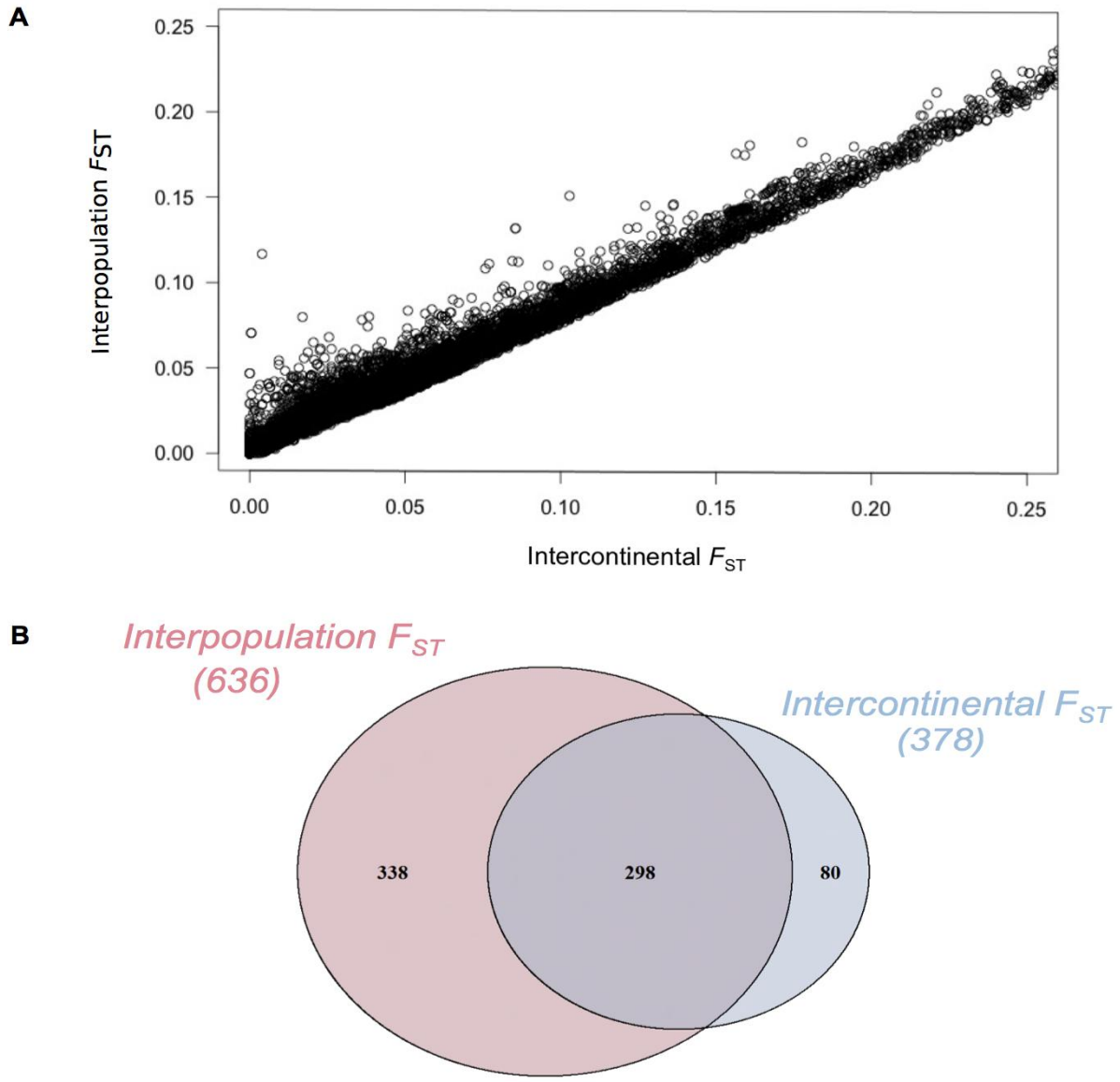


Figure S4. Relationship between interpopulation and intercontinental F_{ST}. (A) Correlation between interpopulation and interpopulation F_{ST} values for the 9695 VIP variants. If slightly lower intercontinental F_{ST} values are observed when compared to interpopulation F_{ST}, a very high correlation between the two indices is observed (Pearson's correlation coefficient = 0.992, *P*-value < 10^{-5}). **(B) Venn diagram depicting the proportions of unique and shared highly differentiated VIP variants across the interpopulation and intercontinental F_{ST} indices.** Among the 9695 VIP variants, 636 (6.6%) and 378 (3.9%) variants have a significant interpopulation F_{ST} and intercontinental F_{ST} value, respectively. A total of 298 variants have significant values for both F_{ST} indices.

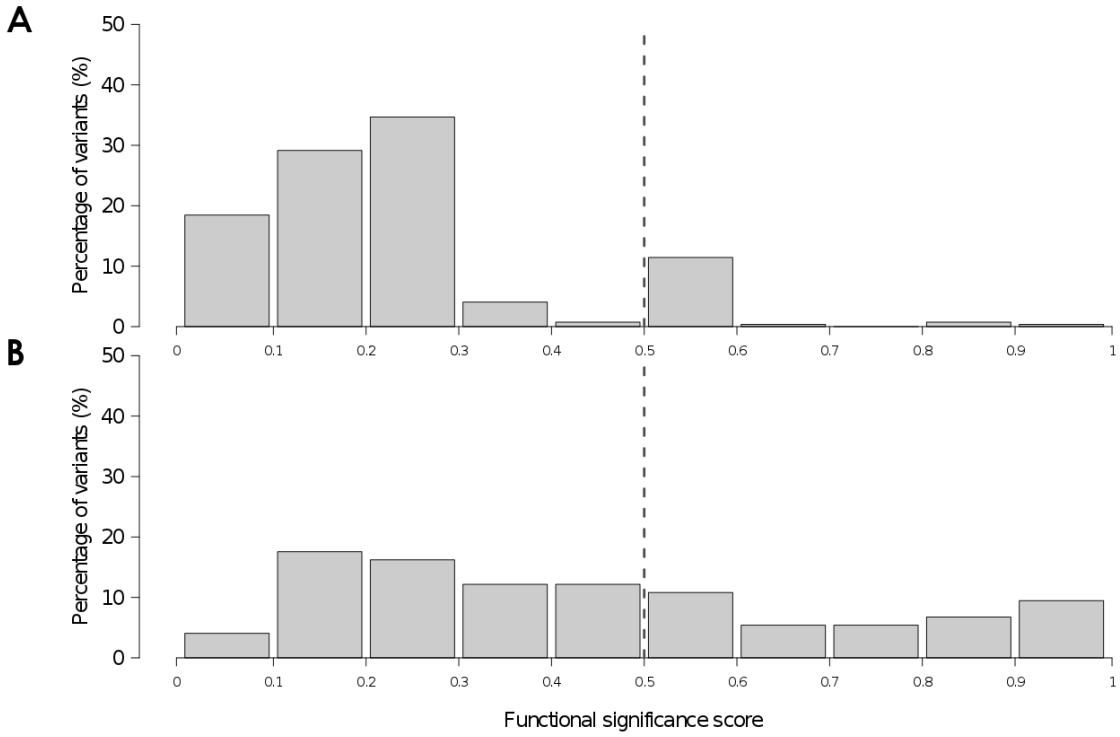


Figure S5. Distribution of functional significance (FS) scores as estimated by the F-SNP tool. Any variant with a FS score ≥ 0.5 (indicated by the dashed gray line) is predicted to have a deleterious effect with respect to one of four major bio-molecular functions (protein coding, splicing regulation, transcriptional regulation and post-translation modifications). The 0.5 threshold was shown to correspond to a false positive rate of 2.2% in a large-scale benchmark dataset of known disease-related SNPs (Lee and Shatkay, 2009). (A) FS scores of the 636 VIP variants with a significant interpopulation F_{ST} value. (B) FS scores of the 90 VIP key variants. Interestingly, the distribution of FS scores for the 90 key variants with a known clinical relevance in pharmacogenetics is significantly different from that of the 636 highly differentiated variants. In particular, the median FS score for the 636 variants is 0.208, whereas, for the key variants, the median rises to 0.400. Moreover, 37.8% of key variants are assigned an FS score ≥ 0.5, whereas only 5.5% of the 636 variants are assigned such a high score. The Kolmogorov–Smirnov test with 5% significance level confirms that the two sets of variants are unlikely to share a common score distribution (P-value = 2.66 x 10^{-11}).

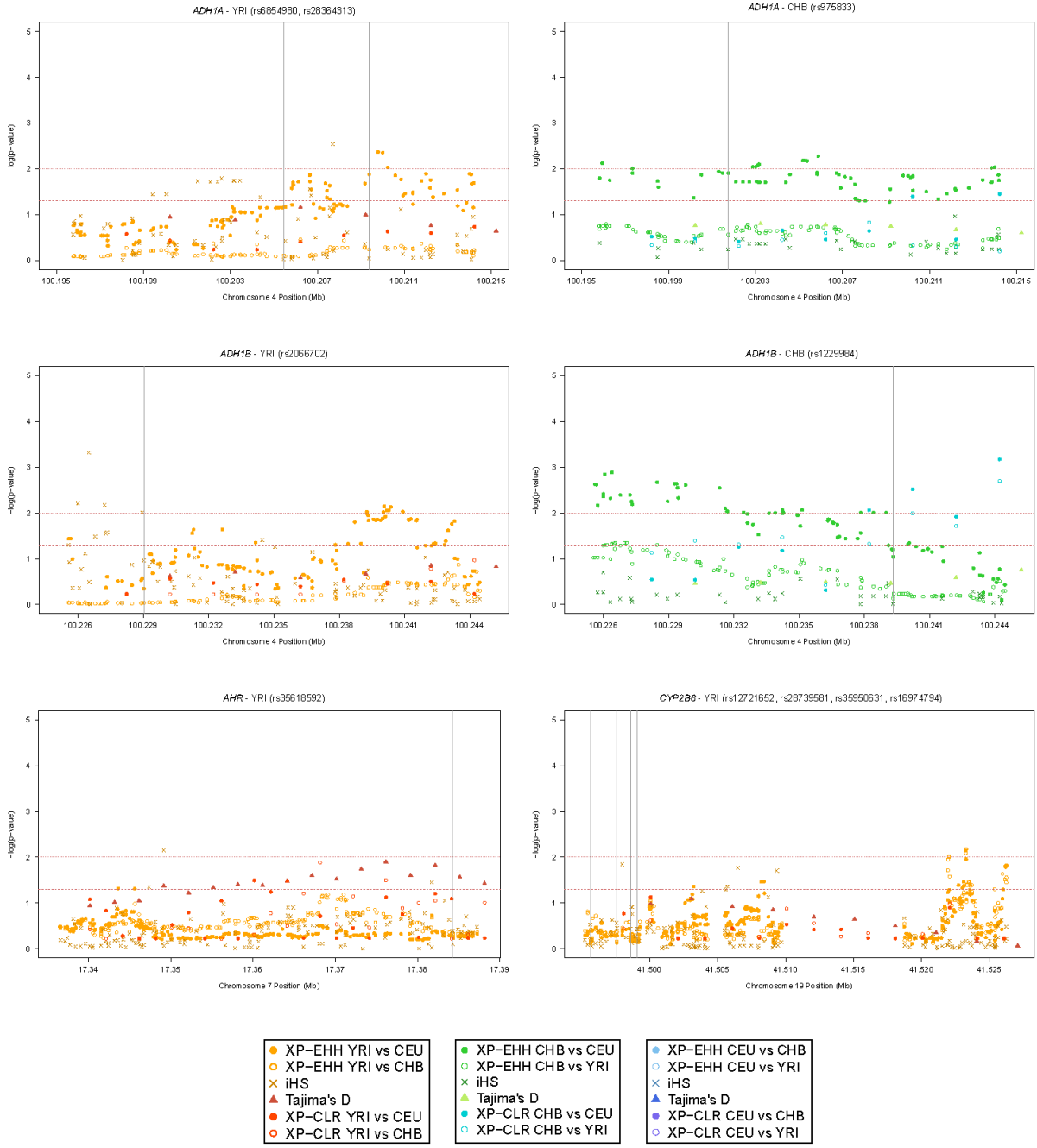


Figure S6. Results of selection tests in the 11 VIP genes where a signal of positive selection was detected in at least one population (YRI, CEU and/or CHB). The distribution of -$\log_{10}$ (*P*-values) are shown for four tests of positive selection: Tajima's *D*, XP-CLR, XP-EHH and iHS. *P*-values are derived from genome-wide empirical distributions of the test statistics, as described in Pybus et al. (2014). Horizontal red dotted and dashed lines show 0.05 and 0.01 genome-wide significance levels, respectively. Black vertical lines indicate the physical position in each gene of either the key variant(s) or the putative functional variant(s) identified by F-SNP (indicated in parentheses above each plot) for which an atypical pattern of geographic differentiation was detected.

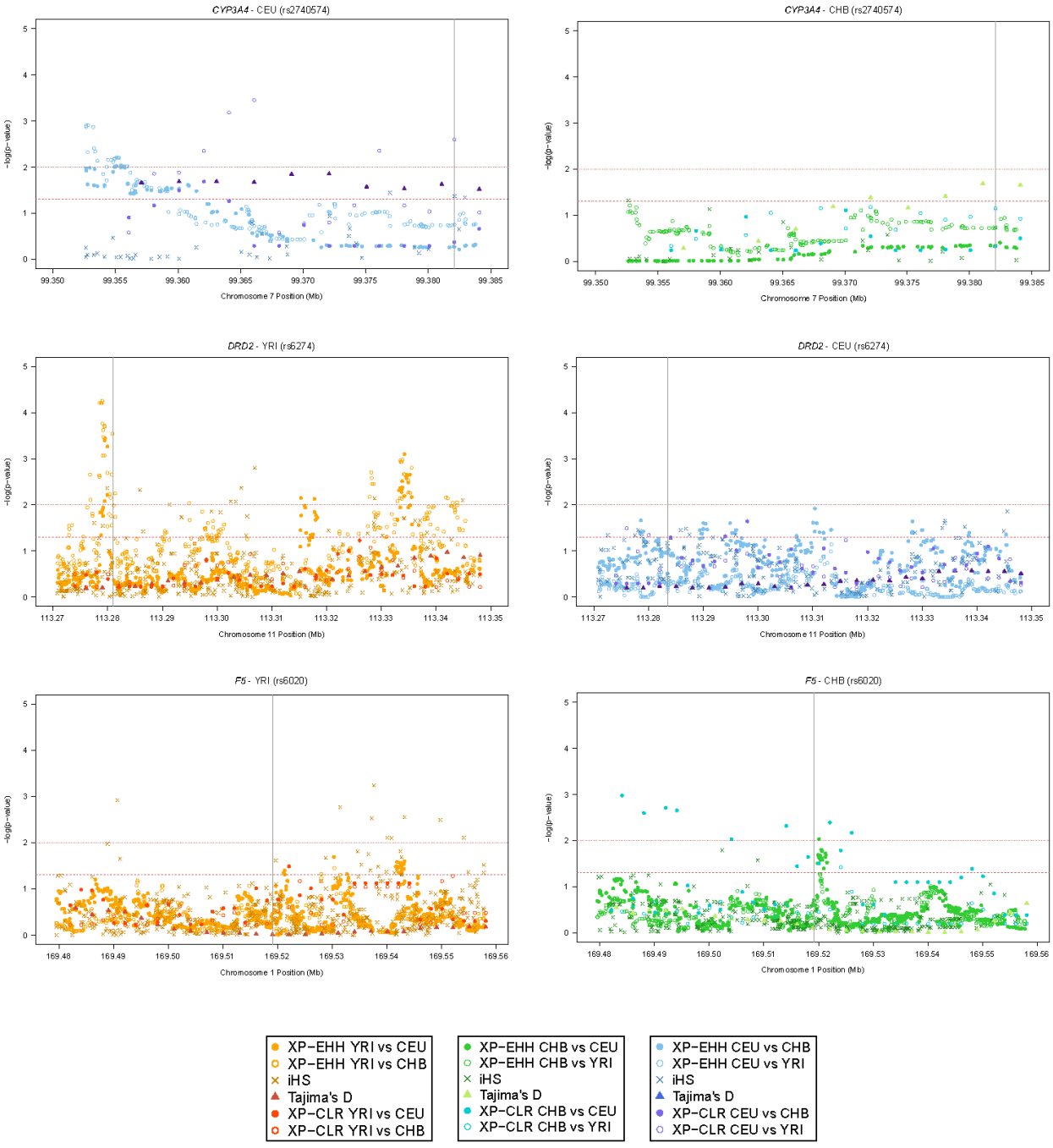


Figure S6. (suite)

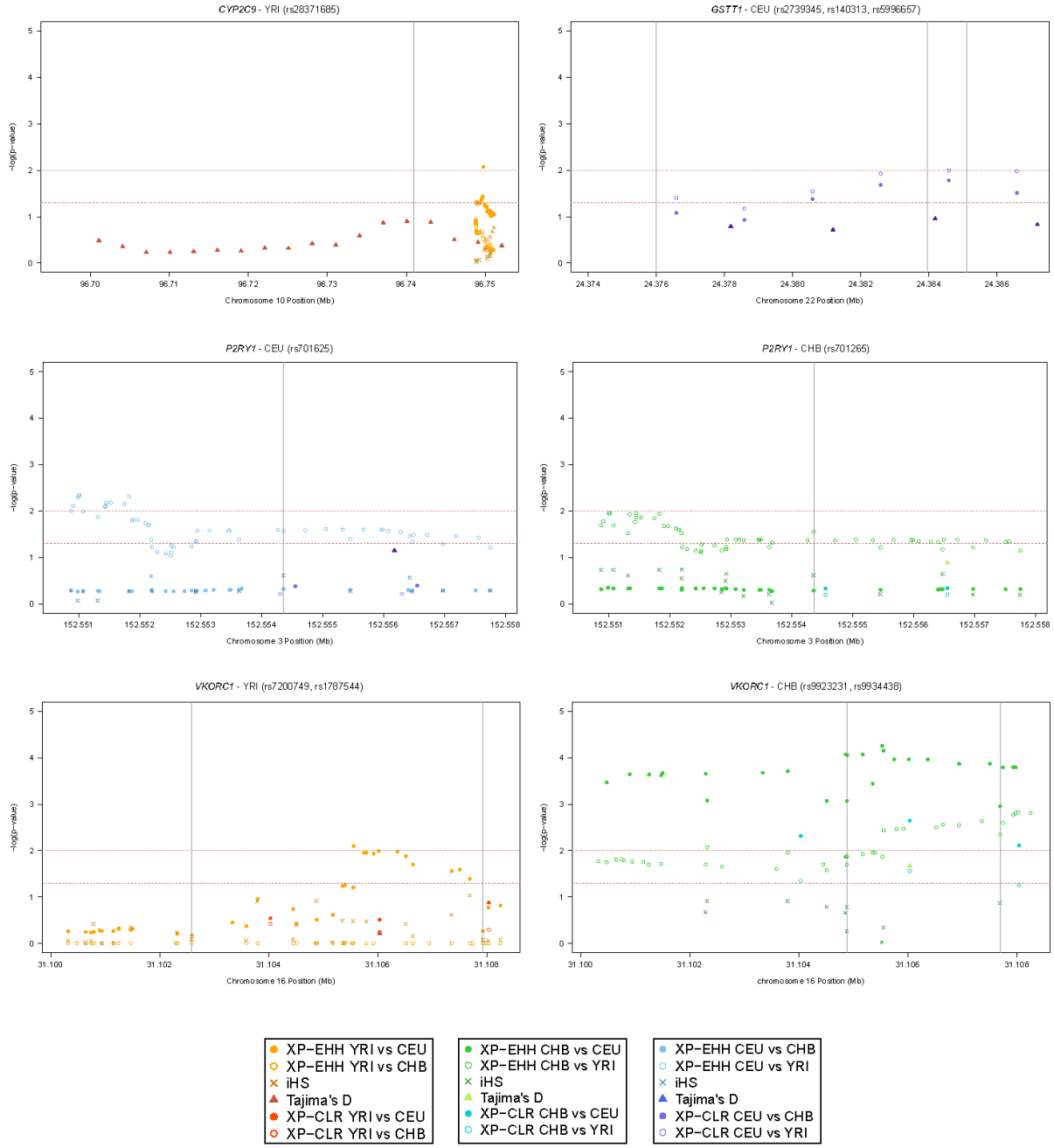


Figure S6. (suite)

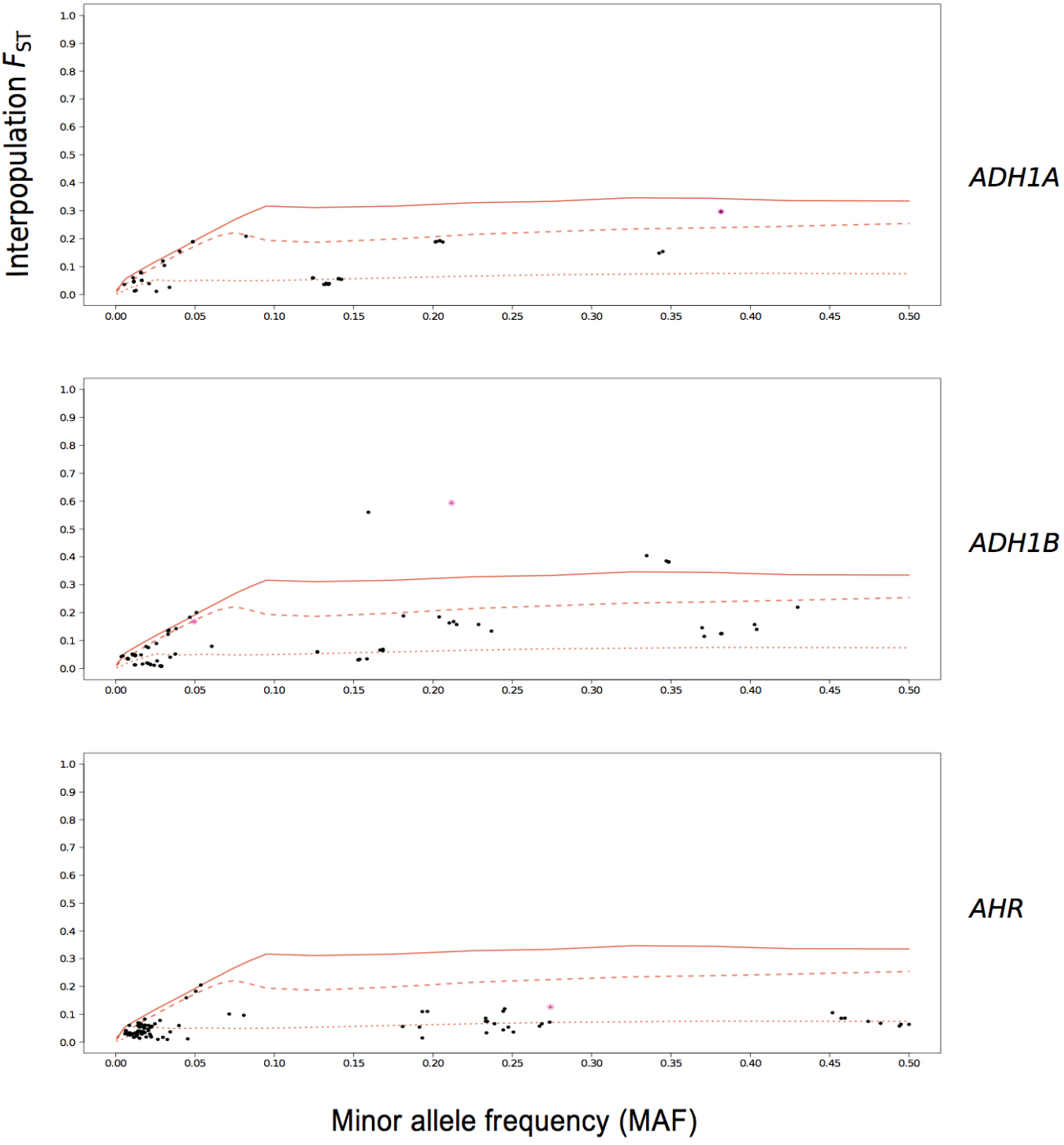


Figure S7. Patterns of genetic differentiation observed for all variants in the eight VIP genes carrying a significant functional variant (FV). Genome-wide empirical distributions of interpopulational F_{ST} values were constructed from 25,532,386 SNVs. The 50th, 95th and 99th percentiles are indicated as dotted, dashed and full red lines, respectively. Individual values of F_{ST} calculated for each variant within the eight VIP genes having a global minor allele frequency (MAF) above 5% in at least one population are plotted against their MAF. FV are shown in pink.

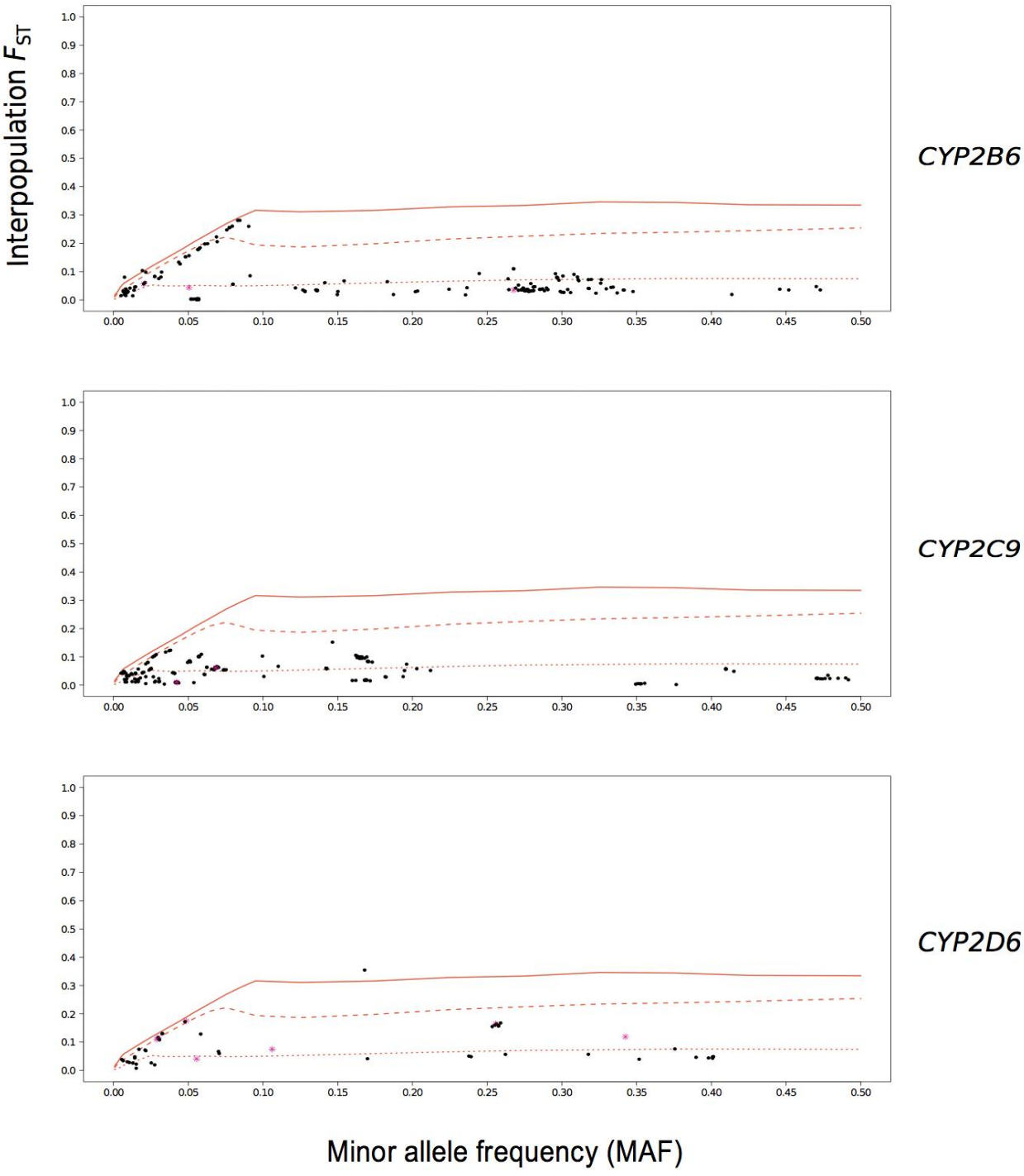


Figure S7. (suite)

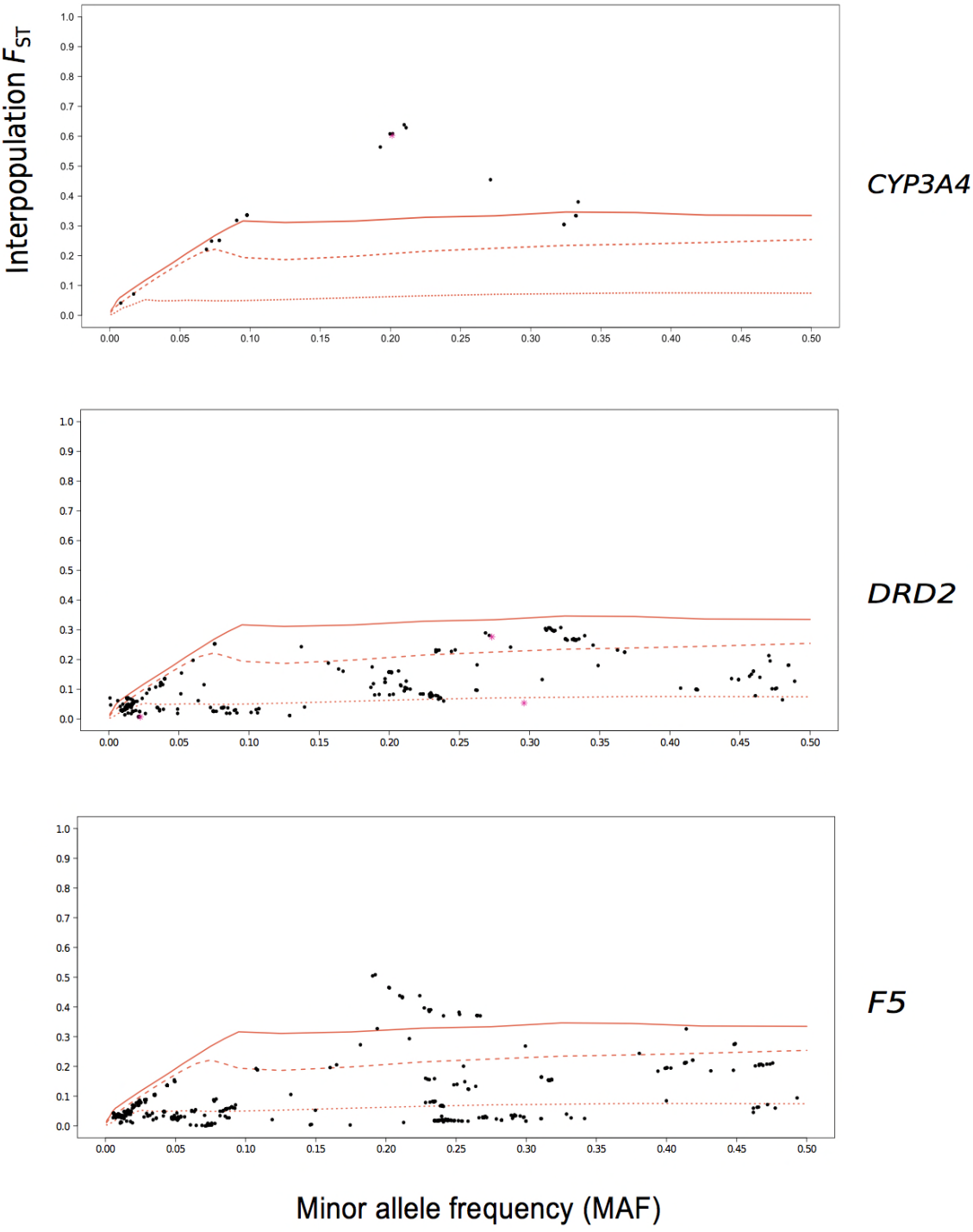


Figure S7. (suite)

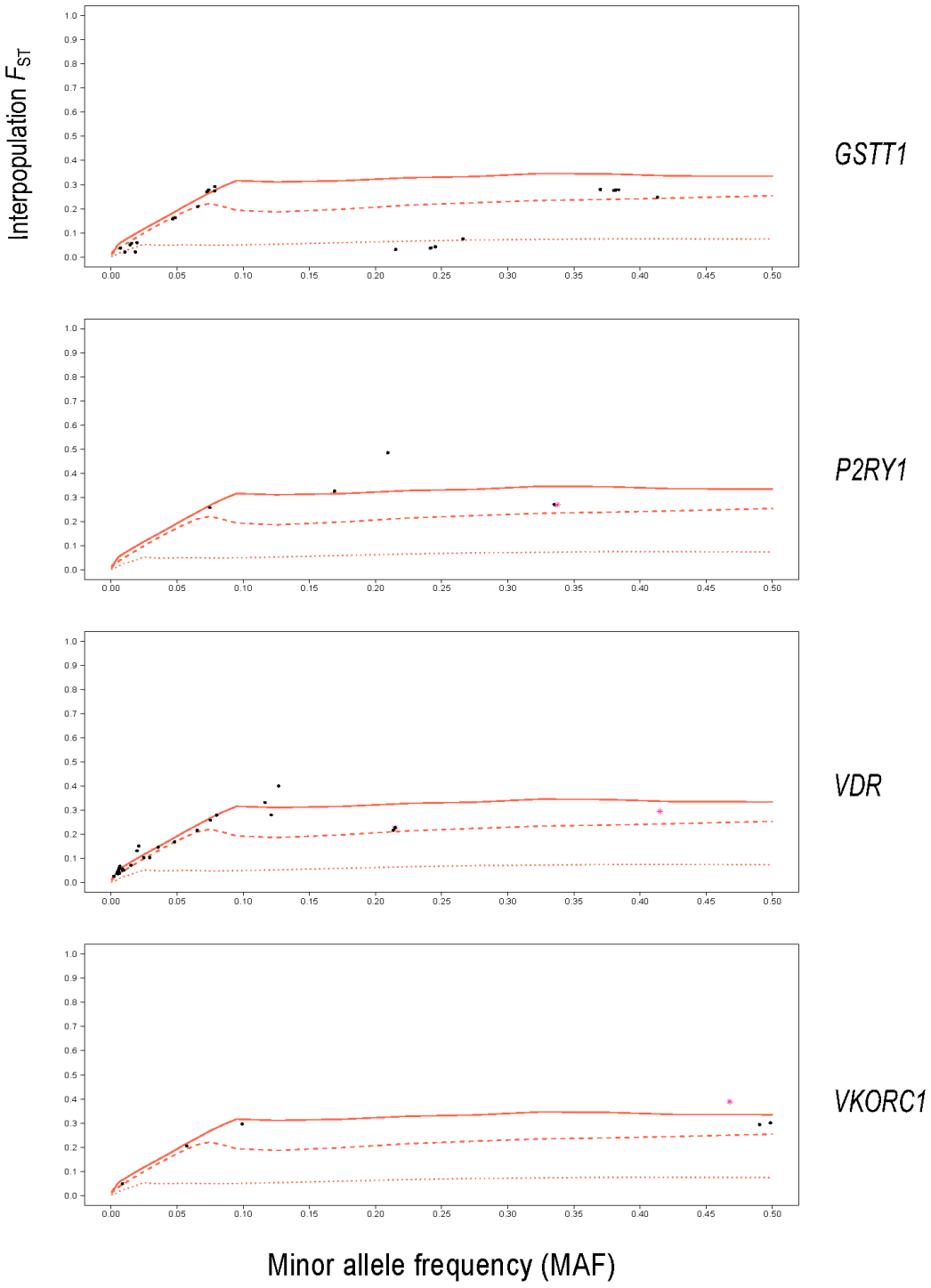


Figure S7. (suite)

Table S1. Description of the 45 VIP genes included in the study.

Gene symbol	Gene name	Physical position[a]	Pharmacogenetic category[b]	# SNVs[c]	# Key variants[d]
ABCB1	ATP-binding cassette, sub-family B (MDR/TAP), member 1	chr7:87131179-87344639	Transporter	601	3
ACE	angiotensin I converting enzyme	chr17:61552422-61577741	Pharmacodynamic	104	0
ADH1A	alcohol dehydrogenase 1A (class I), alpha polypeptide	chr4:100195523-100214185	Phase I metabolism	49	1
ADH1B	alcohol dehydrogenase IB (class I), beta polypeptide	chr4:100225527-100244572	Phase I metabolism	74	2
ADH1C	alcohol dehydrogenase 1C (class I), gamma polypeptide	chr4:100255649-100275917	Phase I metabolism	143	1
ADRB1	adrenergic, beta-1-, receptor	chr10:115801806-115808667	Pharmacodynamic	15	2
ADRB2	adrenergic, beta-2-, receptor, surface	chr5:148204156-148210197	Pharmacodynamic	25	2
AHR	aryl hydrocarbon receptor	chr7:17336276-17387775	Modifier	93	1
ALDH1A1	aldehyde dehydrogenase 1 family, member A1	chr9:75513578-75570233	Phase I metabolism	174	0
ALOX5	arachidonate 5-lipoxygenase	chr10:45867629-45943563	Pharmacodynamic	418	1
BRCA1	breast cancer 1, early onset	chr17:41194312-41279500	Pharmacodynamic	215	0
COMT	catechol-O-methyltransferase	chr22:19927263-19959498	Pharmacodynamic	144	1
CYP1A2	cytochrome P450, family 1, subfamily A, polypeptide 2	chr15:75039184-75050941	Phase I metabolism	23	2
CYP2A6	cytochrome P450, family 2, subfamily A, polypeptide 6	chr19:41347443-41358352	Phase I metabolism	65	5
CYP2B6	cytochrome P450, family 2, subfamily B, polypeptide 6	chr19:41495204-41526301	Phase I metabolism	163	3
CYP2C9	cytochrome P450, family 2, subfamily C, polypeptide 9	chr10:96520463-96614671	Phase I metabolism	271	2
CYP2C19	cytochrome P450, family 2, subfamily C, polypeptide 19	chr10:96696415-96751148	Phase I metabolism	367	3
CYP2D6	cytochrome P450, family 2, subfamily D, polypeptide 6	chr22:42520501-42528883	Phase I metabolism	53	7
CYP2J2	cytochrome P450, family 2, subfamily J, polypeptide 2	chr1:60356980-60394423	Phase I metabolism	108	1
CYP3A4	cytochrome P450, family 3, subfamily A, polypeptide 4	chr7:99352583-99383811	Phase I metabolism	57	1
CYP3A5	cytochrome P450, family 3, subfamily A, polypeptide 5	chr7:99243813-99279621	Phase I metabolism	77	1
DPYD	dihydropyrimidine dehydrogenase	chr1:97541300-98388615	Phase I metabolism	2825	0
DRD2	dopamine receptor D2	chr11:113278317-113348001	Pharmacodynamic	252	3
F5	coagulation factor V	chr1:169479192-169557769	Pharmacodynamic	459	0
GSTP1	glutathione S-transferase pi	chr11:67349066-67356124	Phase II metabolism	46	2
GSTT1	glutathione S-transferase theta 1	chr22:24374139-24386284	Phase II metabolism	25	0
HMGCR	3-hydroxy-3-methylglutaryl-Coenzyme A reductase	chr5:74630993-74659926	Pharmacodynamic	81	3
KCNH2	potassium voltage-gated channel, subfamily H (eag-related), member 2	chr7:150640044-150677402	Pharmacodynamic	125	4
KCNJ11	potassium inwardly-rectifying channel, subfamily J, member 11	chr11:17404796-17412206	Modifier	30	0
MTHFR	5,10-methylenetetrahydrofolate reductase (NADPH)	chr1:11843787-11868160	Pharmacodynamic	96	2
NQO1	NAD(P)H dehydrogenase, quinone 1	chr16:69741304-69762533	Pharmacodynamic	72	1
NR1I2	nuclear receptor subfamily 1, group I, member 2	chr3:119499557-119539332	Modifier	172	0
P2RY1	purinergic receptor P2Y, G-protein coupled, 1	chr3:152550736-152557843	Pharmacodynamic	32	2
P2RY12	purinergic receptor P2Y, G-protein coupled, 12	chr3:151053376-151060585	Pharmacodynamic	17	1
PTGIS	prostaglandin I2 (prostacyclin) synthase	chr20:48118411-48186707	Pharmacodynamic	263	1
PTGS2	prostaglandin-endoperoxide synthase 2 (prostaglandin G/H synthase and cyclooxygenase)	chr1:186638944-186651559	Pharmacodynamic	40	3
SCN5A	sodium channel, voltage-gated, type V, alpha (long QT syndrome 3)	chr3:38587553-38676850	Transporter	375	3
SLC19A1	solute carrier family 19 (folate transporter), member 1	chr21:46932629-46964385	Pharmacodynamic	140	5
SLCO1B1	solute carrier organic anion transporter family, member 1B1	chr12:21282128-21394730	Transporter	743	3
SULT1A1	sulfotransferase family, cytosolic, 1A, phenol-preferring, member 1	chr16:28614908-28622649	Phase II metabolism	60	2
TPMT	thiopurine S-methyltransferase	chr6:18126545-18157374	Phase II metabolism	134	2
TYMS	thymidylate synthetase	chr18:655604-675499	Pharmacodynamic	71	0
UGT1A1	UDP glucuronosyltransferase 1 family, polypeptide A1	chr2:234666919-234683951	Phase II metabolism	58	1
VDR	vitamin D (1,25- dihydroxyvitamin D3) receptor	chr12:48233320-48300814	Pharmacodynamic	320	10
VKORC1	vitamin K epoxide reductase complex, subunit 1	chr16:31100175-31108276	Pharmacodynamic	20	3

Abbreviation: SNV,single nucleotide variant.

[a]Genomic positions according to the GRCh37/hg19 assembly, including 2-kb flanking sequences on either side of each gene

[b]Pharmacokinetic and pharmacodynamic genes are shown in green and orange, respectively.

[c]Total number of SNVs within each gene identified in the 1000 Genomes project that occur with a minor allele frequency > 0.05 in at least one population and which have a functional annotation in dbSNP.

[d]Variants annotated in the VIP summary of the PharmGKB database as key variants with a known clinical relevance in pharmacogenetics.

Table S2. Summary statistics of the nine empirical distributions of the interpopulation F_{ST} statistic built from the genome-wide variation data of the 1000 Genomes project.

		Interpopulation F_{ST}					
Empirical distribution	**Number of SNVs[a]**	**Mean**	**SD**	**Minimum**	**Maximum**	**95th percentile**	**99th percentile**
Genome-wide	25,532,386	0.017	0.036	0.000	0.884	0.080	0.181
Intergenic	15,141,160	0.017	0.036	0.000	0.811	0.081	0.181
Genic	11,282,100	0.017	0.036	0.000	0.811	0.080	0.183
Intronic	10,477,050	0.017	0.036	0.000	0.791	0.080	0.183
5'-UTR	24,395	0.029	0.036	0.000	0.547	0.123	0.231
3'-UTR	198,718	0.020	0.040	0.000	0.667	0.092	0.202
Coding synonymous	107,644	0.024	0.044	0.000	0.705	0.110	0.219
Coding nonsynonymous	146,572	0.017	0.037	0.000	0.811	0.080	0.188
Splice site	1,912	0.016	0.037	0.000	0.400	0.084	0.181

Abbreviations: UTR,untranslated region; SD, standard deviation ;SNV, single nucleotide variant.

[a]Number of SNVs remaining after the application of a LD-based pruning procedure.

Table S3. Interpopulation and intercontinental F_{ST} values for the 9695 variants located in the 45 VIP genes[10]

[10] Fourni sur un tableur excel.